Spring 2005

Bioinformatics study of mammalian MRNA polydenylation

Haibo Zhang
New Jersey Institute of Technology

ABSTRACT

BIOINFORMATICS STUDY OF MAMMALIAN MRNA POLYADENYLATION

by
Haibo Zhang

Messenger RNA polyadenylation is one of the key post-transcriptional events in mammalian cells, which have influences on many aspects of mRNA metabolism. Several human diseases have been shown to associate with abnormal polyadenylation, highlighting the importance of this process. The availability of the complete sequence of human and mouse genomes, together with their gene expression data provides valuable resources to study mRNA polyadenylation on a system level. This dissertation addresses the following issues of mammalian mRNA polyadenylation through bioinformatics approaches: (1) the extensive documentation of several key aspects of polyadenylation sites in humans and mice on a genomic level; (2) the evaluation of whether polyadenylation is an evolutionarily conserved cellular process between human and mouse ortholog genes; (3) the tissue-specificity of the regulation of alternative polyadenylation in humans and mice; (4) the development of a novel approach to use SAGE data to study polyadenylation.

A database is built to comprehensively document mappings of poly(A) sites in humans and mice genome-wide. About 54% of human genes and 32% of mouse genes are shown that can undergo alternative polyadenylation. Conservation studies show that polyadenylation configurations are highly conserved in humans and mice. In addition, Gene Ontology studies identified certain functional groups of genes associated with different types of polyadenylation configurations. Furthermore, tissue-specific usage of

alternative poly(A) sites is observed. Microarray data analysis and sequence analysis identified certain *trans*-acting factors and *cis*-regulatory elements that might be responsible for such regulation of alternative polyadenylation. Finally, a novel approach to use SAGE data to study alternative polyadenylation is developed and demonstrated. The results presented provide important insights into the mechanism of mRNA polyadenylation and a genomic view of the regulation of gene expression by alternative polyadenylation in mammals.

BIOINFORMATICS STUDY OF MAMMALIAN MRNA POLYADENYLATION

by
Haibo Zhang

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey - Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Biology

Federated Biological Sciences Department

May 2005

APPROVAL PAGE

BIOINFORMATICS STUDY OF MAMMALIAN MRNA POLYADENYLATION

Haibo Zhang

Dr. Michael Recce, Dissertation Advisor — Date
Associate Professor, Department of Information Systems, NJIT

Dr. Bin Tian, Dissertation Co-Advisor — Date
Assistant Professor, Department of Biochemistry and Molecular Biology
UMDNJ-New Jersey Medical School

Dr. Wilma Friedman, Dissertation Co-advisor — Date
Assistant Professor, Federated Department of Biological Sciences
Rutgers-Newark

Dr. Wonsuk Yoo, Committee Member — Date
Assistant Professor, Department of Mathematical Sciences, NJIT

Dr. Samuel Gunderson, Committee Member — Date
Associate Professor, Department of Molecular Biology and Biochemistry
Rutgers University

BIOGRAPHICAL SKETCH

Author: Haibo Zhang

Degree: Doctor of Philosophy

Date: May 2005

Undergraduate and Graduate Education

- Doctor of Philosophy in Biology
  New Jersey Institute of Technology and Rutgers, The State University of New Jersey – Newark, Newark, NJ, 2005

- Master of Science in Molecular Microbiology
  Nankai University, Tianjin, P.R. CHINA, 1999

- Bachelor of Science in Microbiology
  Nankai University, Tianjin, P.R. CHINA, 1996

Major: Computational Biology

Presentations and Publications:

Zhang, H., Recce, M., and Tian, B. (2005) Tissue-specific alternative polyadenylation in humans. Manuscript in preparation.

Zhang, H., Recce, M., and Tian, B. (2005) A SAGE view of human alternative polyadenylation. Manuscript in preparation.

Pan, Z., Zhang, H., Lutz, C.S., and Tian, B. (2005) An intronic poly(A) site in human and mouse CstF-77 genes suggests an evolutionarily conserved regulatory mechanism. RNA. Submitted.

Hall-Polgar, T., Zhang, H., Tian, B., and Lutz, C.S. (2005) Alternative polyadenylation of COX-2: a mechanism for gene regulation, Nucleic Acids Research 2005 May 4;33(8):2565-79.

Zhang, H., Hu, J., Recce, M., and Tian, B. (2005) PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Research 2005 Jan 1;33 Database Issue:D116-20.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Research. 2005 Jan 12;33(1):201-212.

Zhang, H., Ramanathan, Y., Soteropoulos, P., Recce, M., and Tolias, P.P. (2002) EZ-Retrieve: A web-server for batch retrieval of coordinate-specified human DNA sequences and underscoring putative transcription factor-binding sites. Nucleic Acids Research 30(21): E121-E121.

Ramanathan, Y., Zhang, H., Aris, V., Soteropoulos, P., Aaronson, S.A., and Tolias, P.P. (2002) Functional Cloning, Sorting and Expression Profiling of Nucleic-Acid Binding Proteins. Genome Research. 12(8): 1175-84

Wang, Y., Zhang, H., Chen, S., Wang, J., Wu, J., Wu, D., You, B., and Sun, J. Isolation, Purification and Analysis of A Polysaccharide From Residues of Glycyrrhiza Uralensis Fisch Acta Scientiarum Naturalium Universitatis Nankaiensis, 1999,32(3): 36-38, article in Chinese.

Wang, J., Peng, J., Zhang, H., and Wang, Y. Research on the fermentative characteristics and physical and chemical property of an antagonistic strain of Bacillus subtilis Bs-98Acta Scientiarum Naturalium Universitatis Nankaiensis, 1996,29(4): 89-94, article in Chinese.

Sodhi, A., Zhang, H., and Recce, M. A Simulation of Biological Descent with Modification. Poster at the 13th Sigma Xi Student Research Symposium at St. Joseph's University.

I would like to dedicate this dissertation to my wife, Po Hu, and my parents, Xueyu Zhang and Huizhen Sun, for having supported me through my education and encouraging me to strive for excellence.

"The world is full of rules and principles, don't fight against them, flow with them!"

My Mom

ACKNOWLEDGMENT

I would like to express my deepest appreciation to Dr. Michael Recce, who not only has been a great mentor, a best friend, and a source of inspirations of novel ideas, but also a constant support during my graduate studies. I would also like to thank Dr. Bin Tian who lets me to work in his lab and has been a great supervisor during the last year of my research. Many thanks to Dr. Peter Tolias for supporting me and directing me while I was doing my cooperative education at the Center for Applied Genomics. Special thanks are given to Dr. Wilma Friedman for actively supervising my dissertation research, and to Dr. Wonsuk Yoo and Dr. Samuel Gunderson for participating in my committee.

I am also very grateful to the Director Dr. Patricia Soteropoulos at the Center for Applied Genomics. I would like to thank Dr. Y. Ramanathan and Dr. Lutz for their help with my research, and other colleagues Dr. Virginie Aris, Jeff Cheng, Michael Cody, Cheng Fan, Anthony Galante, Saleena Ghanny, Lisa Hague, Tyra Hall-Polgar, Jun Hu, Songchun Liang, Jianghui Liu, Zenhua Pan, Dr. Vasilis Papasotiropoulos, Filipo Posta, Anbing Shi, Qi Wang, Tongsheng Wang, and Donna Wilson.

I would like to thank Dr. Ronald Kane, Karen Gansner, Amy Trimarco and Clarisa González-Lenahan for helping me with the "administrative" side of this doctorate.

I thank my sister, Dr. Haiying Zhang, and my brother, Dr. Hailong Zhang, for their love and support throughout my education.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Mammalian mRNA Polyadenylation

Almost all mammalian proteins-encoding mRNAs go through several mRNA processing reactions before they become mature. These include capping, splicing, and polyadenylation. Polyadenylation of mammalian mRNAs is a two-step process, which includes a specific cleavage at the 3' end of a nascent mRNA and an addition of a poly(A) tail (Colgan and Manley, 1997). Both reactions are directed by *cis*-regulatory elements and complicated protein machineries. A large number of mammalian genes have more than one polyadenylation site, the choice of which can be regulated by alternative polyadenylation (AP). Alternative polyadenylation together with alternative splicing (AS) and alternative initiation are major mechanisms that contribute to the large-amount of transcriptomic pool in mammalian species (Proudfoot et al., 2002).

Polyadenylation is a key step in mRNA 3'-end formation and has impact on many aspects of mRNA metabolism in the cell, including mRNA stability, mRNA localization, and translation (Lewis et al., 1995). The correct regulation of polyadenylation is crucial in normal cell growth and development (Lou et al., 1996; Peterson et al., 1991; Takagaki and Manley, 1998). Aberrant polyadenylation can lead to various human diseases such as hereditary thrombophilia, metachromatic leukodystrophy, thalassaemia, and amyotrophic lateral sclerosis (ALS), highlighting the importance of the regulation of polyadenylation (Gehring et al., 2001; Gieselmann et al., 1989; Higgs et al., 1983; Lin et al., 1998; Orkin et al., 1985). Research on the mechanism of regulation of mammalian polyadenylation is

important to a better understanding of the process and will provide potential solutions to certain human diseases.

Besides nuclear polyadenylation, mammalian cells also have cytoplasmic polyadenylation during oocyte maturation and early development (Richter, 1999). The research of this dissertation is focused solely on nuclear polyadenylation. In addition, it is worth pointing out that replication-dependent histone genes terminate at the 3' end without a poly(A) tail, instead a stem-loop structure is formed (Dominski and Marzluff, 1999). It has also been reported that there are polyA$^-$ mRNA pools found primarily in the brain that are different from polyA$^+$ transcripts (Brilliant et al., 1984; Van Ness et al., 1979). However, whether these are really non-polyadenylated transcripts or just artifacts are still unclear (Fung et al., 1991; Snider and Morrison-Bogorad, 1992).

1.2 Regulation of Mammalian mRNA Polyadenylation

Both *cis*-regulatory elements and *trans*-acting factors are involved in the regulation of polyadenylation. In mammalian systems, several mRNA sequence elements and many protein factors have been identified. The biochemistry of some of them has also been well defined, whereas others are under rigorous investigation in recent years.

1.2.1 Sequence Determinants on mRNA

The most well-defined and highly conserved polyadenylation signal (PAS) is AAUAAA located ~10-35 nt upstream of the polyadenylation cleavage site. AAUAAA is originally identified from a comparative study of mRNAs from human, rabbit, mouse, and chicken (Proudfoot and Brownlee, 1976). Since then, the essential role of this hexanucleotide in

polyadenylation has been established by extensive mutagenesis studies. From sequenced polyadenylated transcripts, it was estimated that AAUAAA exists in at least 90% of them, and most of other 10% are AUUAAA (Colgan and Manley, 1997). With the availability of human genomic sequences and expressed sequence tag (EST) data, large-scale bioinformatics studies have discovered that AAUAAA only exist in 60-70% of polyadenylated mRNAs (Gautheret et al., 1998; Graber et al., 1999b; Legendre and Gautheret, 2003). Besides AUUAAA, ten single-base variants of AAUAAA also were found to have a significant occurrence rate (Beaudoing et al., 2000). It is generally believed that mRNA with these variants is associated with tissue-specific alternative polyadenylation (MacDonald and Redondo, 2002).

Further upstream of the PAS is the upstream elements (USE). USE has been mainly studied in viral poly(A) site such as simian virus 40 (SV40) late and human immunodeficiency virus type 1 (HIV-1) (Carswell and Alwine, 1989; Schek et al., 1992; Valsamakis et al., 1992). Mammalian genes also have such elements (Brackenridge et al., 1997; Brackenridge and Proudfoot, 2000; Moreira et al., 1998; Moreira et al., 1995) and the existence of USE are likely to be a common feature of polyadenylation signals (Legendre and Gautheret, 2003). USE can regulate polyadenylation by increase the cleavage efficiency and the processitivity of poly(A) addition (Brackenridge and Proudfoot, 2000). USE can be recognized by U1A, PTB, and a recently identified hFip1p (Chan and Black, 1997; Kaufmann et al., 2004; Phillips and Gunderson, 2003). Downstream of polyadenylation cleavage sites is the U- or G/U-rich elements (DSE), which are also usually required for efficient polyadenylation. The position of DSE to the downstream of poly(A) sites are important in such regulation (Chen and Wilusz, 1998;

Chou et al., 1994; Gil and Proudfoot, 1987; McDevitt et al., 1986; McDevitt et al., 1984). U- or G/U-rich DSE has also been identified by large-scale multiple eukaryotic species study (Graber et al., 1999b). A recent study also suggest that sequences between downstream GU-rich elements can be bound by U1A and negatively regulate polyadenylation (Phillips et al., 2004).

Besides PAS, USE, and DSE, local sequence compositions surrounding the cleavage site are also important. Large-scale sequence analysis showed that the nucleotide composition in ±100nt of a cleavage site is dramatically different from those further upstream or downstream (Graber et al., 1999b; Legendre and Gautheret, 2003). Mutagenesis evidence suggests that there are preference to nucleotide at the cleavage position, often is a CA dinucleotide (Chen et al., 1995). Figure 1.1 upper panel shows a schematic overview of the anatomy of sequence elements around a poly(A) site.

In addition to nucleotide compositions and sequence motifs, mRNA secondary structure may play roles in the regulation process of polyadenylation. In virus, some secondary structures seem important in keeping AAUAAA open for binding and hence positively regulate the usage of specific poly(A) sites (Guntaka, 1993), while other sequences downstream of AAUAAA maybe also important in forming secondary structures (Hans and Alwine, 2000). In humans, such regulations by stem-loop structures also exists (Phillips et al., 1999; Phillips et al., 2004).

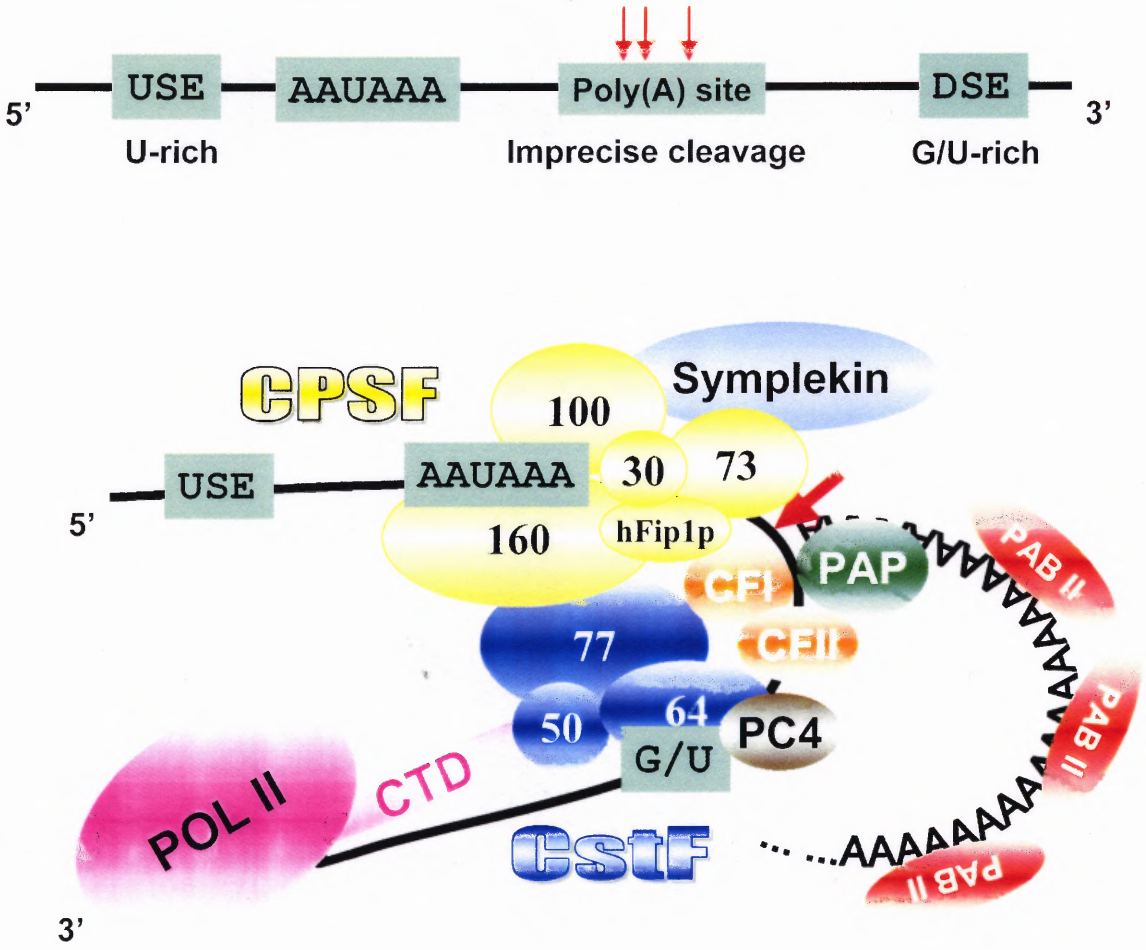


Figure 1.1 Schematic representation of sequence determinants around polyadenylation sites and polyadenylation machineries in mammalian system. Upper panel shows sequence determinants and imprecise cleavage sites (see chapter 1.3). USE: upstream sequence elements; DSE: downstream sequence elements. Lower panel shows protein factors. CPSF: cleavage and polyadenylation specific factor; CstF: cleavage stimulatory factor; CTD: Pol II C-terminal domain; PAP: poly(A) polymerase; PAB II: poly(A) binding protein. Numbers correspond to molecular weight of CPSF and CstF subunits. CPSF 160, 100, 73 and 30 kDa subunits are ortholog of Yeast Yhh1p, Ydh1p, Ysh1p and Yth1p respectively; CstF 77 and 64 kDa subunits are ortholog of Yeast Rna14p and Rna15p respectively; hFip1p is ortholog of Yeast Fip1p; PC4 is ortholog of Yeast Sub1p. Some factors that may be involved are not shown, see text for details.

1.2.2 Cleavage and Polyadenylation Machinery

There are multiple protein factors required for the full function of the cleavage and the adding of the poly(A) tail. The polyadenylation machineries are generally conserved between mammal and yeast. While most factors have been subjected to extensive studies in yeast, looking for ortholog counterparts in mammalian systems have recently helped in identifying more protein factors with regulatory roles in polyadenylation (Calvo and Manley, 2001; Kaufmann et al., 2004). In mammals, core protein factors required for cleavage and polyadenylation include cleavage/polyadenylation specificity factor (CPSF), cleavage stimulatory factor (CstF), cleavage factors I and II (CFI and CFII), RNA polymerase II (Pol II, mainly C-terminal Domain, CTD), poly(A) polymerase (PAP), and poly(A) binding protein II (PABII) (Figure 1.1). However, there is still no definitive evidence for the long-sought endonuclease. The most probable ones are CPSF-73 and CPSF-30 (Ryan et al., 2004; Zarudnaya et al., 2002).

CPSF is required for both the cleavage and poly(A) addition reactions. CPSF complex was originally purified from cattle and humans containing four subunits of molecular weight about 160kD, 100kD, 73kD, 30kD (Bienroth et al., 1991; Murthy and Manley, 1992). Recently, a fifth subunit hFip1p of about 66kD was identified in human HeLa cells (Kaufmann et al., 2004). CPSF-160 recognizes the AAUAAA signal and directly cross-links to RNA at AAUAAA and USE positions (Gilmartin et al., 1995; Moore et al., 1988). CPSF-160 also interacts with CstF-77, PAP, and hFip1p (Kaufmann et al., 2004; Murthy and Manley, 1995). In vitro studies show hFip1p can stimulate PAP activity in a USE or U-rich region dependent manner, this is most likely by recruiting PAP to RNA through its arginine-rich RNA-binding (ARM) motif at the C-terminus

(Kaufmann et al., 2004). CPSF-100 shares high sequence similarities with CPSF-73, but their functions are not well defined. A pair of paralogs named RC-68 (homolog of CPSF-73) and RC-74 (homolog of CPSF-100) was recently shown to also form a complex, however, neither of the paralogs was found in CPSF complex. Depletion of RC-68 by RNAi arrests cells in G1 phase but doesn't affect cell growth (Dominski et al., 2005). The yeast homolog of CPSF-73 contains typical motifs of metallo-β-lactamase. Mutagenesis studies support yeast CPSF-73 homolog having a putative hydrolytic lactamase domain, the activity of which is Zn(2+) dependent. The facts that both cleavage and polyadenylation are supported by Zn(2+) and that CPSF-73 contacts the cleavage site in an AAUAAA-dependent manner suggest that CPSF-73 is the long-sought endonuclease, although *in vitro* proof of endonuclease activity of purified CPSF-73 is still lacking (Ryan et al., 2004). CPSF-30 is thought to cooperate with CPSF-160 and with PABP II. Although it has RNA binding properties, CPSF-30 is not always detected in active CPSF (Murthy and Manley, 1992). *Drosophila* homolog of CPSF-30, CLP, has endoribonuleolytic activity and has been demonstrated to bind G/C-rich clusters with Zinc-finger domains, which are also Zn(2+) dependent (Bai and Tolias, 1996; Bai and Tolias, 1998). However, because purified CPSF without CPSF-30 can function for cleavage *in vitro* (Murthy and Manley, 1992), it seems unlikely that CPSF-30 is the endonulease for the cleavage reaction.

Unlike CPSF, CstF is only required for the cleavage step. CstF consists of three subunits of 77kD, 64kD, and 50kD. CstF-77 bridges between CstF-64 and CstF-50 and directly interacts with CPSF-160, probably being mediated by the HAT (half a TPR) domains, which are also exist in its *Drosophila* and yeast homologs (Murthy and Manley,

1995; Takagaki and Manley, 1994). CstF-64 binds to DSE GU-rich region (Perez Canadillas and Varani, 2003) and this binding is coordinated by the interaction between CPSF-160 and CstF-77 in an AAUAAA dependent fashion (MacDonald et al., 1994; Takagaki and Manley, 1997). CstF-50 is similar to mammalian G protein beta subunits and contains WD-40 motifs (Takagaki and Manley, 1992). It can bind to both phosphorylated and unphosphorylated Pol II CTD, however, not through the WD-40 motifs, but through the N-terminal 95 amino acids. Over-expression of CstF-50 CTD-binding domain can inhibit the cleavage reaction in a dominant-negative and dose-dependent fashion, but without a disruption the CstF complex (Fong and Bentley, 2001). Taken together, CstF binds pre-mRNA and interacts with Pol II CTD and CPSF to form the core of cleavage/polyadenylation machinery.

Besides CPSF and CstF, there are two cleavage factors CF I and CF II that are required for cleavage (Takagaki et al., 1989). In 2000, de Varies *et al* partially purified CF II and showed that it is a two-subunit complex, one of which (CFIIAm or hClp1) interacts with CPSF-160 and CF I (de Vries et al., 2000). In addition, Pol II CTD is also required for cleavage and interacts with CPSF and CstF, coupling 3' end processing with transcription (Hirose and Manley, 1998; Hirose and Manley, 2000; Hirose et al., 1999; McCracken et al., 1997). A heptad YSPTSPS is repeated 52 times in human Pol II CTD and 26 times in yeast ortholog. *In vitro* data suggests there are differences between different parts and maybe merely length dependent (Fong and Bentley, 2001; Ryan et al., 2002). Phosphorylation and dephosphorylation at serine 2 and serine 5 of Pol II CTD are critical in the regulation of capping, splicing, and 3' end formation (Hirose and Manley, 2000; Proudfoot, 2000). Finally, the addition of poly(A) tail is carried out by PAP, which

also interacts with CPSF. The newly synthesized poly(A) tail is bound by PABP II and this in turn increased PAP processitivity (Wahle and Ruegsegger, 1999).

In addition to these core protein complexes, recent molecular dissections of individual cases have identified several more protein factors that play regulatory roles in the polyadenylation process, including Symplekin/Pta1 (Takagaki and Manley, 2000; Xing et al., 2004), PC4/Sub1 (Calvo and Manley, 2001), Ssu72 (He et al., 2003), hnRNP F (Veraldi et al., 2001), hnRP H/H' (Arhin et al., 2002), U2AF65 (Millevoi et al., 2002), U1A (Gunderson et al., 1994; Gunderson et al., 1998; Gunderson et al., 1997; Lutz and Alwine, 1994; Lutz et al., 1996), PTB (Castelo-Branco et al., 2004), and SRp20 (Lou et al., 1998). The fact that most of these newly identified factors are also involved in mRNA splicing machinery support the current view of the tight coupling among transcription, splicing, and polyadenylation (Calvo and Manley, 2003; Edwalds-Gilbert et al., 1997; Orphanides and Reinberg, 2002; Proudfoot et al., 2002). In fact, it has also been shown that the general transcription factors (GTF) TFIID interacts with CPSF (Dantonel et al., 1997) during transcription initiation. A complete list of all factors is summarized in Table 1.1.

Table 1.1 Polyadenylation-related factors.

Official Symbol	Protein	Gene ID
POLR2A	RNA polymerase II polypeptide A, 220kDa, C-Terminal Domain (CTD)	5430
POLR2B	RNA polymerase II polypeptide B, 140kDa	5431
POLR2C	RNA polymerase II polypeptide C, 33kDa	5432
POLR2D	RNA polymerase II polypeptide D	5433
POLR2E	RNA polymerase II polypeptide E, 25kDa	5434
POLR2F	RNA polymerase II polypeptide F	5435
POLR2G	RNA polymerase II polypeptide G	5436
POLR2H	RNA polymerase II polypeptide H	5437
POLR2I	RNA polymerase II polypeptide I	5438
POLR2J	RNA polymerase II polypeptide J	5439
POLR2K	RNA polymerase II polypeptide K	5440
POLR2L	RNA polymerase II polypeptide L	5441
CPSF1	Cleavage and polyadenylation specificity factor 1, 160kDa	29894
CPSF2	Cleavage and polyadenylation specificity factor 2, 100kDa	53981
CPSF3	Cleavage and polyadenylation specificity factor 3, 73kDa	51692
CPSF4	Cleavage and polyadenylation specificity factor 4, 30kDa	10898
CPSF5	CFIM, cleavage factor Im, 25kDa	11051
CPSF6	CFIM, cleavage factor Im, 68kDa	11052
CSTF1	Cleavage stimulatory factor subunit 1, 50kDa	1477
CSTF2	Cleavage stimulatory factor subunit 2, 64kDa	1478
CSTF3	Cleavage stimulatory factor subunit 3, 77kDa	1479
CSTF2T	Cleavage stimulatory factor subunit 2, 64kDa, tau variant	23283
HEAB	ATP/GTP binding protein, component of CFIIAm(de Vries et al., 2000; Licatalosi et al., 2002)	10978
PCF11	Pre-mRNA cleavage complex II protein(de Vries et al., 2000)	51585
PAPOLA	Poly(A) polymerase alpha	10914
PAPOLB	Poly(A) polymerase beta (testis specific)	56903
PAPOLG	Poly(A) polymerase gamma	64895
PABPC1	Poly(A) binding protein, cytoplasmic 1	26986
PABPN1	Poly(A) binding protein, nuclear 1	8106
PABPC3	Poly(A) binding protein, cytoplasmic 3	5042
PABPC4	Poly(A) binding protein, cytoplasmic 4	8761
SYMPK	Symplekin(Takagaki and Manley, 2000; Xing et al., 2004)	8189
HNRPF	Heterogeneous nuclear ribonucleoprotein F(Veraldi et al., 2001)	3185
HNRPH1	Heterogeneous nuclear ribonucleoprotein H1 (H)(Zarudnaya et al., 2003)	3187
HNRPH2	Heterogeneous nuclear ribonucleoprotein H2 (H')(Zarudnaya et al., 2003)	3188
U2AF2	U2 (RNU2) small nuclear RNA auxiliary factor2, U2AF65(Millevoi et al., 2002)	11338
SNRPA	U1 small nuclear ribonucleoprotein polypeptide A(Gunderson et al., 1994; Gunderson et al., 1997; Lutz and Alwine, 1994; Lutz et al., 1996)	6626
PC4	Transcriptional coactivator PC4(Calvo and Manley, 2001; Ge and Roeder, 1994)	10923
LOC286528	Ssu72, only demonstrated in yeast(He et al., 2003)	286528
SFRS3	SRp20(Lou et al., 1998)	6428
PTB1P	Polypyrimidine tract binding protein(Castelo-Branco et al., 2004)	5725

1. All factors are described by review articles (Edwalds-Gilbert et al., 1997; Proudfoot, 1996; Proudfoot et al., 2002; Zhao et al., 1999) if not otherwise noted.

1.3 Mammalian Alternative Polyadenylation

In mammalian systems, a gene can have only one poly(A) site or more than one poly(A) site. Genes with more than one poly(A) site can be subject to the regulation of alternative polyadenylation, through which transcripts with different 3'UTR or different protein products can be obtained from the same gene. Alternative polyadenylation thus greatly contributes to the diversity of transcriptome.

The most well studied cases of alternative polyadenylation include regulation of IgM heavy chain mRNA in B cell differentiation (Peterson et al., 1991; Phillips et al., 2001; Takagaki and Manley, 1998) and regulation of calcitonin/calcitonin gene-related peptide (CGRP) (Lou et al., 1996), which have been extensively reviewed (Edwalds-Gilbert et al., 1997; Proudfoot, 1996; Zhao et al., 1999). Edwalds-Gilbert *et al.* have compiled ~120 genes known to be subjected to alternative polyadenylation, ~30 of them show evidence of coupling with alternative splicing (Edwalds-Gilbert et al., 1997). The general theme of regulation lies in two aspects: the first is that the components of processing complex CPSF and CstF can be regulated and respond differently to the sequence determinants around a poly(A) site; the second is that further factors can act to repress or activate the processing components, either by competing or enhancing the binding to sequence determinants or via protein-protein interactions.

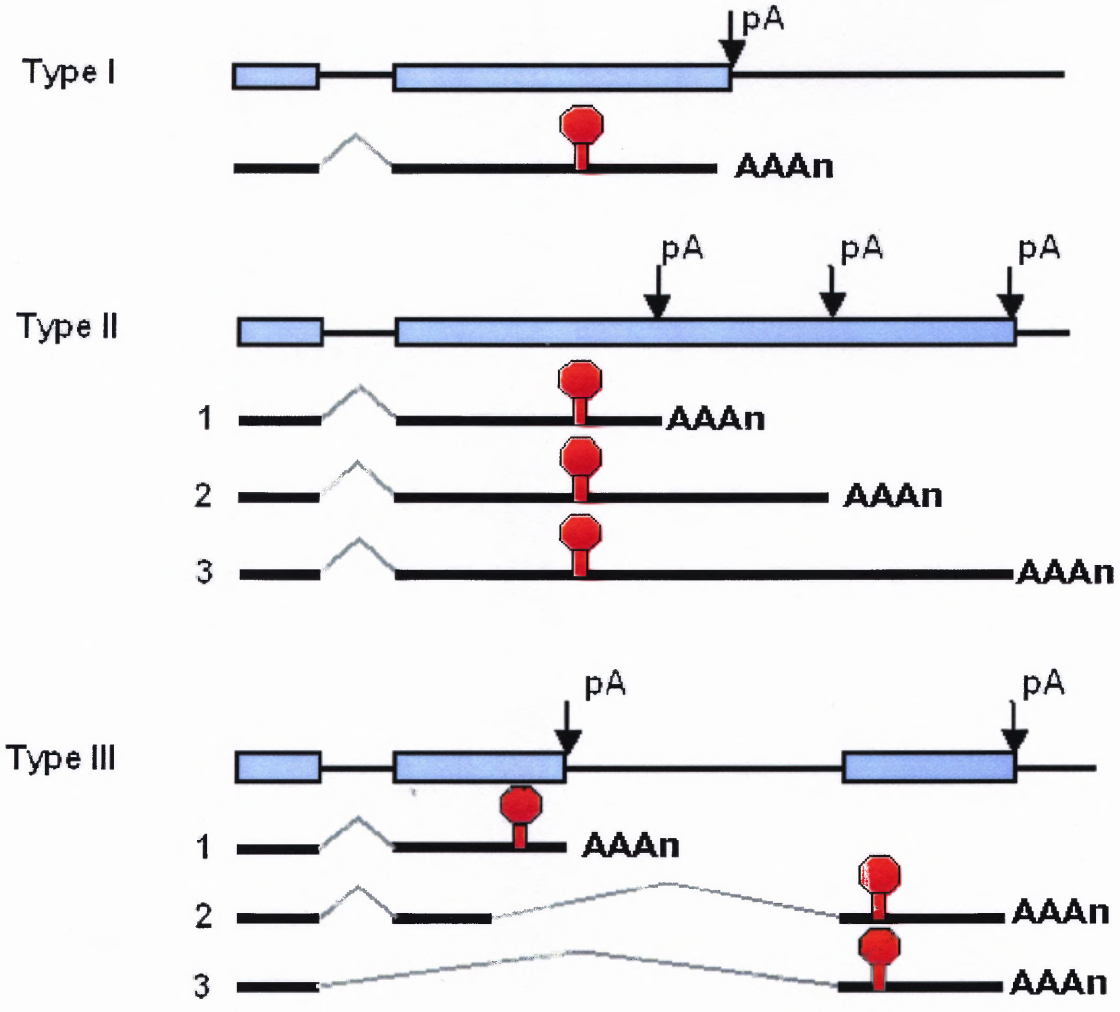


Figure 1.2 Schematic representation of polyadenylation sites in different types of genes. Stop codon is indicated in red; pA stands for poly(A) sites. Only 3' of a gene is shown.

Based on number of poly(A) sites and their locations relative to a gene structure, a gene can be classified into one of the three types (Figure 1.2): Type I: genes with only one poly(A) site; Type II: genes with more than one poly(A) site, all locating in the 3' most exon; and Type III: genes with more than one poly(A) sites, locating on different exons. These three types will be referred as polyadenylation configurations throughout the dissertation. A large-scale study has estimated that 29% of mammalian genes can

undergo alternative polyadenylation (Type II and Type III polyadenylation configurations). It is generally believed that variants of sequence determinants are associated with weak poly(A) sites and that the 3' most alternative poly(A) sites are stronger than those 5' proximal ones (Beaudoing et al., 2000). The usage of alternate poly(A) sites has been indicated to be tissue- or development-specific (Edwalds-Gilbert et al., 1997). However, the exact mechanism for how one poly(A) site is chosen over another under different physiological conditions still needs in depth studies. The understanding of the regulation of such an important cellular process on a genomic level should have great impact on our knowledge of mRNA metabolism, genome structure and function, and certain diseases.

A poly(A) site can be determined by the interaction of protein factors and sequence determinants upstream and downstream of the cleavage site. However, upon the assembly of polyadenylation machineries, the actual cleavage site could be imprecise (Chen et al., 1995; Natalizio et al., 2002; Sheets et al., 1990). This type of alternate polyadenylation cleavage sites is different from alternative poly(A) sites resulted from using different sequence determinants. This is referred to as imprecise or heterogeneous cleavage of the polyadenylation process. As a result, a poly(A) site refers to a region that imprecise cleavage sites might occur (Figure 1.1 upper panel). *In silico* study and SAGE data suggest that the region is about 17nt in size (Beaudoing and Gautheret, 2001; Pauws et al., 2001).

1.4 Bioinformatics Studies of Alternative Polyadenylation

In the past two decades, there were great advances in the fields of genome sequencing projects and the large-scale gene expression profiling, such as expressed sequence tags (EST). To date, human genome is nearly complete and the EST database (dbEST) at National Center for Biotechnology Information (NCBI) contains over six million human ESTs and four million mouse ESTs (Boguski et al., 1993) (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). In addition, NCBI gene expression omnibus (GEO) provides huge amount of large-scale gene expression data (http://www.ncbi.nlm.nih.gov/geo/) (Edgar et al., 2002). These resources provide unprecedented opportunities for us to study mammalian alternative polyadenylation systematically.

1.4.1 Expressed Sequence Tags

An EST is a segment of a cDNA sequence reverse transcribed from an mRNA. It could be from either the 5' or 3' of a full or partial cDNA clone. Because most ESTs carry mature transcript information, EST has been a great resource for the systematic large-scale studies of alternative splicing (Modrek et al., 2001; Xu and Lee, 2003; Xu et al., 2002). On the same line, when aligned to genomic sequences, poly(A) sites can be readily identified by the end positions of the aligned ESTs resulted from the cleavage reaction of polyadenylation (Figure 1.3). Because aligned EST counts reflect transcript levels and cDNA library sources provide tissue, developmental, and cancer/disease information, ESTs provide vast amount of information to study many aspects of polyadenylation.

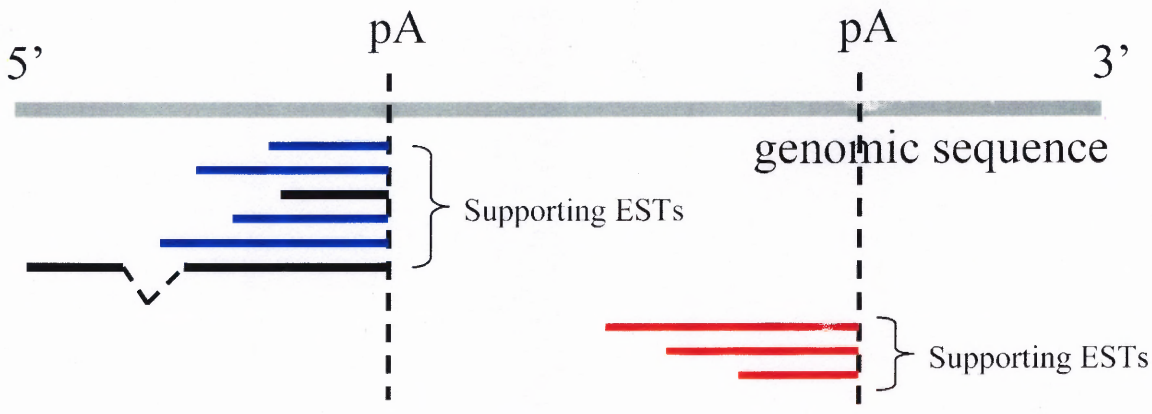


Figure 1.3 Aligning ESTs to genomic sequences can identify poly(A) sites. Solid gray line represents genomic sequence; ESTs colors represent different tissue resource of cDNA libraries; pA stands for poly(A) site.

The first large-scale study of polyadenylation using ESTs were carried out in 1998 (Gautheret et al., 1998). 189 human EST clusters showed clear evidence supporting alternative polyadenylation. The low number of discovery is mostly because of the limited number of ESTs available at the time (164,000). Also because of this, all ESTs are used to infer poly(A) sites by examining the end positions of EST-to-genome alignments. However, this could be problematic. Because an EST is a partial sequence from an mRNA, the end of an alignment does not necessarily represent a genuine poly(A) site. To overcome this problem, one can use only those ESTs containing poly(A)/(T) tails, providing a clear evidence of the end of the pre-mRNA and the addition of the poly(A) tail. Even so, another possible artifact, internal priming, also needs to be checked. Internal priming could happen during the synthesis of cDNA using oligo(dT) primers when an internal poly(A) stretch is present on an mRNA, resulting in poly(A)/(T) tailed ESTs that do not reflect real poly(A) sites. Most of these internal

priming events can be identified by inspecting the genomic sequences for the existence of a poly(A) stretch (Figure 1.4).

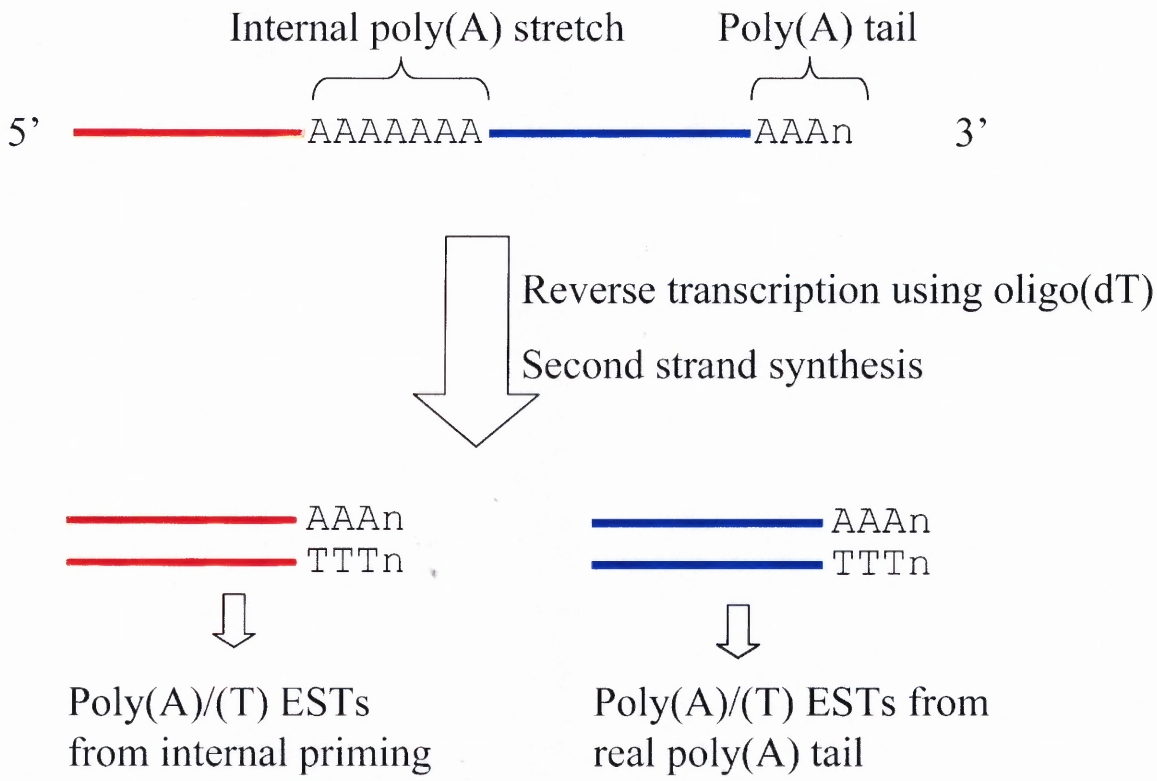


Figure 1.4 ESTs from internal priming events do not reflect real poly(A) sites.

1.4.2 Other Large-Scale Data Resources

Although only EST has so far been demonstrated to be applicable for the large-scale bioinformatics study of polyadenylation, there are several other types of resources having the potential to facilitate the better understanding of this process (Table 1.2). These include Microarray data, Serial Analysis of Gene Expression (SAGE) data, and Gene Ontology. As different sources of large-scale data are poised to address different aspects

of functional genomics, cross-platform integrations of these data will provide a more comprehensive view of polyadenylation and compensate each other for their limitations.

Table 1.2 Large-scale data resources.

Resource	Description and URL	Citation
NCBI dbEST	NCBI database of Expressed Sequence Tags: http://www.ncbi.nlm.nih.gov/dbEST/	(Boguski et al., 1993)
Gene Ontology	Gene Ontology Categorized by biological process, cellular components, and molecular functions: http://www.geneontology.org/	(Ashburner et al., 2000)
SAGE Net	SAGE data and tag mappings http://www.sagenet.org/	
NCBI GEO	NCBI Gene Expression Omnibus, contains microarray, SAGE, and other types of expression data: http://www.ncbi.nlm.nih.gov/geo/	(Edgar et al., 2002)
EBI ArrayExpress	European Bioinformatics Institute ArrayExpress public microarray database: http://www.ebi.ac.uk/arrayexpress/	(Brazma et al., 2003)
SMD	Stanford Micoarray Database: http://genome-www.stanford.edu/microarray	(Ball et al., 2005)

1.4.2.1 Micorarray Data

The well-established microarray platform is the Affymetrix gene chip. A gene chip is a high-density nucleotide array contains thousands of probesets representing different genes. Each probeset is formed by several oligonucleotides (often 16), which are called probes and encompass different regions of the gene's mRNA. By hybridizing fluorescence labeled mRNA samples of interest, messenger levels can then be assessed simultaneously for thousands of genes (Lockhart et al., 1996). Vast amount of microarray data are available in various public domains such as NCBI GEO (Edgar et al., 2002), the Stanford Microarray Database (SMD) (Ball et al., 2005), and ArrayExpress at European Bioinformatics Institute (EBI) (Brazma et al., 2003) (Table 1.2).

The potential of using microarray data to study polyadenylation lies in two aspects: 1) to apply computational techniques on the probe level to dissect effects of alternative polyadenylation; 2) to apply transcriptional profiling techniques to study regulated *trans*-acting protein factors across different physiological conditions.

Because there are 16 probes encompassing different regions of an mRNA, some of them might encompass alternative exons. Therefore, affymetrix probes can be used to study the relative expression level of splicing variants resulting from alternative splicing (Lockhart and Winzeler, 2000; Xu and Lee, 2003; Xu et al., 2002). On the same line, probes encompass regions of mRNA that could be differentially expressed because of the choice of different poly(A) sites can be also used to study alternative polyadenylation. However, because current Affymetrix probes are not specifically designed for such purpose, only a small portion of genes can be studied through this method. However, microarray data in public domains (Table 1.2) contains transcription profiles of thousands of genes across various physiological conditions. Combining transcriptional profiling of polyadenylation factors using microarray data with EST data can therefore facilitate the *trans*-factor study of polyadenylation on a large-scale.

1.4.2.2 SAGE Data

Serial Analysis of Gene Expression (SAGE) was initially developed to detect and quantify the expression of large numbers of transcripts (Velculescu et al., 1995). In a SAGE experiment, all poly(A) tailed transcripts are first reverse transcribed into cDNA sequences. Using an anchoring enzyme (AE, usually NlaIII) and a tagging enzyme (TE, typically BsmFI), a short sequence tag (10-14bp) can be obtained to represent each

transcript. Because the tag is obtained from a unique position within each transcript, it can uniquely identify the corresponding transcript. Sequence tags for all transcripts are then linked together to from long serial molecules that can be cloned and sequenced. This promotes the efficiency of sequencing to multiple folds in terms of transcript discovery. When counting the number of times a particular tag is observed, it actually quantifies the expression level of the corresponding transcript. Therefore, SAGE is currently mainly applied in the field of gene expression profiling across different physiological conditions (Figure 1.3). There are two advantages of SAGE compared to EST and Microarray: 1) SAGE data is much more quantitative than EST data in measuring transcription levels; 2) because SAGE tags are intrinsically more biased to 3' end of transcripts than microarray probes, more transcripts maybe distinguished by distinct SAGE tags. Because polyadenylation is an event happens primarily at the 3' of a transcript, it is expected that methods could be developed to take advantage of SAGE data to study polyadenylation.

1.4.2.3 Gene Ontology

The Gene Ontology (GO) is a project with the goal to consistently describe the function of all gene products. GO terms are controlled vocabularies that are organized through hierarchical tree structures in three principles: biological process, cellular components, and molecular functions. Genes and gene products are being associated with different GO terms based on various information resources such as biological databases and literature publications (Ashburner et al., 2000). When a group of genes are being studied, important functional groups can be readily revealed by analysis of GO terms that these genes are associated with. As discussed before, three types of genes exist in mammalian system

based on number of poly(A) sites and their positions, GO provides a means to study whether such types of polyadenylation configurations have any functional implications.

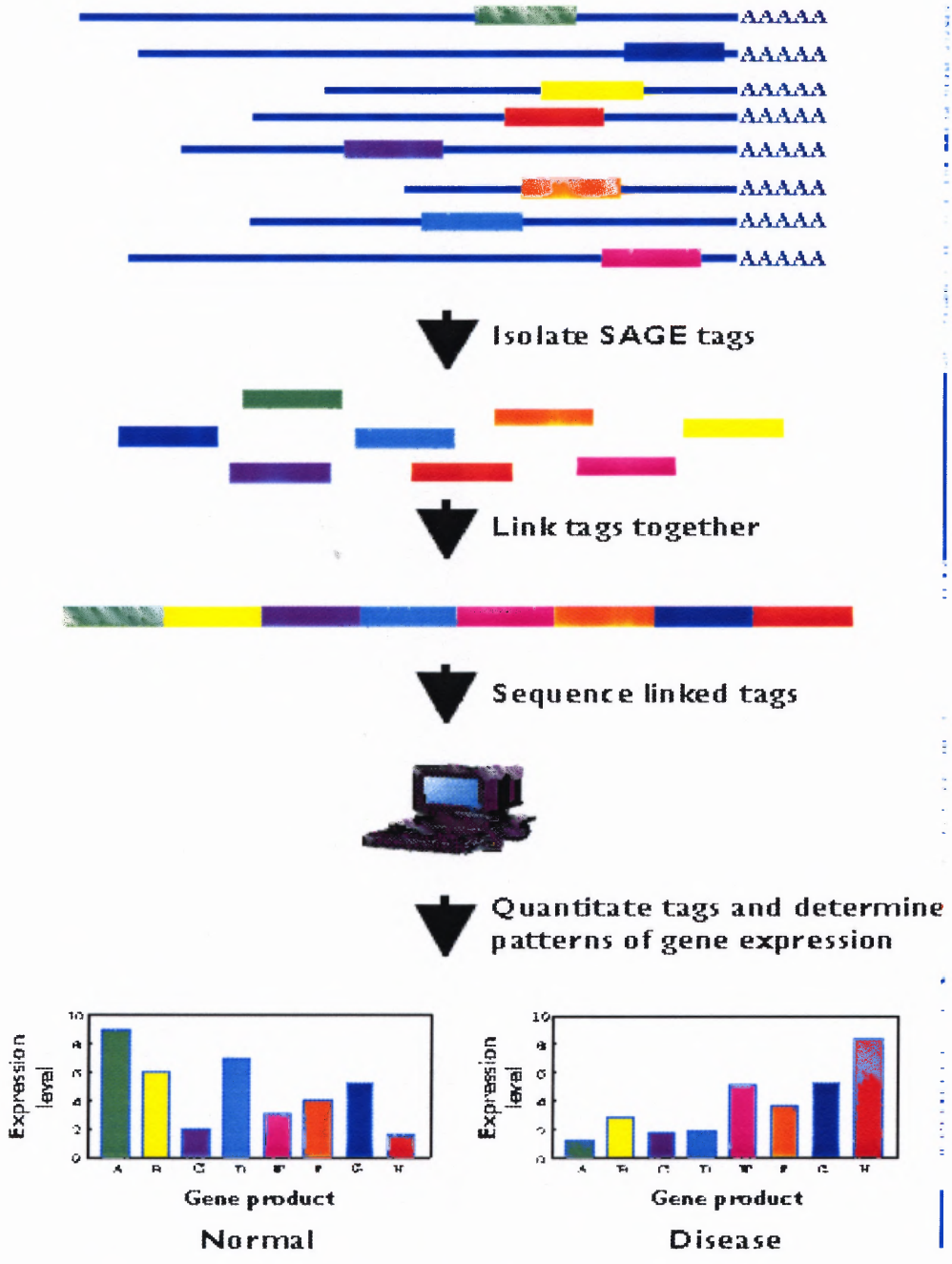


Figure 1.5 Schematic illustration of SAGE method. Adapted from SAGE net (http://www.sagenet.org/) website.

1.4.3 Challenges in Large-Scale Bioinformatics Analysis of Polyadenylation

With the accumulation of EST data, more alternate poly(A) sites have been discovered and large-scale studies of sequence determinants and tissue-specificities are conducted (Beaudoing et al., 2000; Beaudoing and Gautheret, 2001; Graber et al., 1999b; Legendre and Gautheret, 2003). However, these studies were mostly UTR-based instead of gene-based, extensive documenting of poly(A) site for human and mouse gene in a genomic context is noticeably lacking. In addition, multi-species study was limited to the comparison and the discovery of sequence determinants around poly(A) sites, whereas conservations in polyadenylation configurations regarding number of poly(A) sites and their positions with respect to gene exon/intron structures have yet to be conducted. Furthermore, with the availability of a much more comprehensive Gene Ontology resource (Ashburner et al., 2000), it is possible to assess on a system level the functional aspects of polyadenylation as a fundamental cellular process. Finally, although tissue-specific alternative polyadenylation has been implicated (Beaudoing and Gautheret, 2001), several key questions of tissue-specific alternative polyadenylation events have not been comprehensively addressed, these include the tissue-specific usage of strong versus weak poly(A) sites, tissue-specific positional preferences, and tissue-specific *trans*-acting factors and *cis*-regulatory elements that might account for the tissue-specificity of alternative polyadenylation events.

There are several known issues in using ESTs to study polyadenylation, notably sequencing errors, internal priming, presence of chimeric ESTs and paralog genes, potential vector contamination at the ends, and inclusion of genomic sequence because of incomplete splicing. From quantification point of view, although EST counts can reflect

relative transcript levels, variations among experimental methods of cDNA libraries introduce much noise when analyzing EST data. In addition, some cDNA libraries are normalized for gene discovery and economic reasons. As a result, EST counts from normalized libraries cannot be used to infer mRNA levels (Bonaldo et al., 1996). Recent years, more and more oligo-nucleotide based microarray and serial analysis of gene expression (SAGE) data are becoming available (Lockhart et al., 1996; Lockhart and Winzeler, 2000; Velculescu et al., 1995) and have been used in studies of alternative splicing (Johnson et al., 2003; Wang et al., 2003). How to take advantage of these huge amounts of expression data to the studies of alternative polyadenylation in mammalian systems calls for new computational approaches.

1.5 Summary and Outline

Mammalian mRNA polyadenylation is an important cellular process that affects many aspects of mRNA metabolism. It involves a cleavage reaction and a poly(A) addition reaction at the 3' end of a pre-mRNA. Sequence motifs around poly(A) sites and many protein factors are important in regulating the process. A large portion of mammalian genes have more than one poly(A) site and may be regulated by alternative polyadenylation. Bioinformatics approaches can be applied to study on a system level the regulation of alternative polyadenylation using the available public EST data. There exist great challenges in large-scale studies of conservation, gene ontology, tissue-specificity, and new computational approaches utilizing other expression data, which will all greatly contribute to our further understanding of polyadenylation on a system level.

The first step toward the system level study of mRNA polyadenylation is to document all polyadenylation sites genome-wide. Next chapter presents the construction of a polyadenylation database toward this effort, which lays the foundation for subsequent systematic analysis of mammalian polyadenylation. Chapter 3 presents the conservation and gene ontology studies. Tissue-specific alternative polyadenylation studies integrating EST data and microarray data are discussed in chapter 4. Finally, a novel approach is demonstrated in chapter 5 to use SAGE data to study alternative polyadenylation.

The importance of studying mRNA polyadenylation lies in many aspects, including the elucidation of mRNA metabolism, the understanding of genome structures and functions, and the cure to certain diseases. While biochemical approaches are poised to study the functions of individual elements or factors, a bioinformatics approach aims at the systematic analysis of alternative polyadenylation and the evaluation of its biological impact. The results of bioinformatics studies presented in this dissertation will greatly facilitate the mechanistic studies of mRNA polyadenylation and have a far-reaching impact on the understanding of the functional dynamics of mammalian genomes.

CHAPTER 2

POLYA_DB: POLYADNEYLATION DATABASE FOR HUMANS AND MICE

2.1 Abstract

Messenger RNA polyadenylation is one of the key post-transcriptional events in eukaryotic cells. A large number of genes in mammalian species can undergo alternative polyadenylation, which leads to mRNAs with variable 3' ends. As the 3' end of mRNAs often contains *cis*-regulatory elements important for mRNA stability, mRNA localization, and translation, the implications of the regulation of polyadenylation can be multifold. Alternative polyadenylation is controlled by *cis*-regulatory elements and *trans*-acting factors, and is believed to occur in a tissue- or disease-specific manner. Given the availability of many databases devoted to other aspects of mRNA metabolism, such as transcriptional initiation and splicing, systematic information on polyadenylation, including alternative polyadenylation and its regulation, is noticeably lacking. A database named PolyA_DB is presented here, through which several types of information regarding polyadenylation in mammalian species are presented: (1) polyadenylation sites and their locations with respect to the genomic structure of genes; (2) *cis*-regulatory elements surrounding polyadenylation sites; (3) comparisons of polyadenylation configuration between ortholog genes; (4) tissue/organ information for alternative polyadenylation sites. Currently PolyA_DB contains 45,565 polyadenylation sites for 25,097 human and mouse genes, representing the most comprehensive polyadenylation database to date. The database is accessible via http://polya.umdnj.edu/polyadb.

2.2 Introduction

It has been estimated that more than 29% of human genes have alternative poly(A) sites (Beaudoing and Gautheret, 2001). The choice of alternative poly(A) sites is believed to be related to biological conditions such as cell types and disease states (Edwalds-Gilbert et al., 1997). Alternative polyadenylation can lead to mRNA with variable 3' ends, or proteins with different C-termini. A growing number of genes have been found to be regulated by this mechanism. However, a public database systematically providing information on alternative polyadenylation is lacking. The availability of genomic sequences from several mammalian species as well as large number of expressed sequence tags (ESTs) makes it feasible to comprehensively document mRNA polyadenylation configurations for genes. While ESTs provide both sequence data and information on the biological origin of transcripts by the means of cDNA library source, they have several problems with respect to data quality, such as chimeric sequence, vector contamination, and inclusion of genomic sequence. In addition, when dealing with polyadenylation, issues such as internal priming and low quality sequences at the 5' and 3' ends are more palpable. Therefore a computational approach to study polyadenylation must take these into consideration to ensure that poly(A) sites are accurately mapped. A computational pipeline has been developed in the lab and applied to human and mouse genes, which effectively utilizes genomic sequences and EST data to study polyadenylation (Tian et al., 2005). To further document all aspects of this information, a database needs to be built for mammalian polyadenylation information.

Several key aspects of polyadenylation should be taken into consideration when constructing a comprehensive polyadenylation database: 1) systematic views should be

provided in a genomic context, where all poly(A) sites should be annotated based on gene structures aligned to genomic sequences; 2) because *cis*-regulatory elements, especially polyadenylation signals, are very important in the regulation process, they should be annotated for all poly(A) sites located; 3) both human and mouse EST dataset are available for large-scale analysis, therefore comparisons of polyadenylation configurations between evolutionary conserved ortholog genes will provide great insight in understanding the mechanisms of regulation and should be documented in the database; 4) EST alignments to genomic sequences are crucial evidence supporting the existence of a poly(A) site and hence should be stored in the database; 5) cDNA libraries of ESTs provide great amount of tissue or disease information, which can be also associated to poly(A) sites by their supporting ESTs. Finally, to make the data documented in the database available for individual researchers, a web-based user-interface is desirable to provide accesses to the aforementioned aspects of mammalian polyadenylation.

Currently, the database documents 45,565 poly(A) sites and various information regarding the sites, including their genome locations, evidence of cDNA/EST alignments to genomes, *cis*-regulatory elements surrounding poly(A) sites, comparison of polyadenylation configuration between orthologs, and tissue/organ information for poly(A) sites. A web site is also built for users to access the database from the World Wide Web (http://polya.umdnj.edu/polyadb). While only human and mouse poly(A) sites are currently documented in the database, the data process pipeline and the structure of the database are designed so as to make it easy to include other species in the future. This resource can be of great value to researchers who are interested in studying the

mechanism of polyadenylation and gene regulation by alternative polyadenylation alike. It lays the foundation of further large-scale analyses of mammalian mRNA polyadenylation in this dissertation.

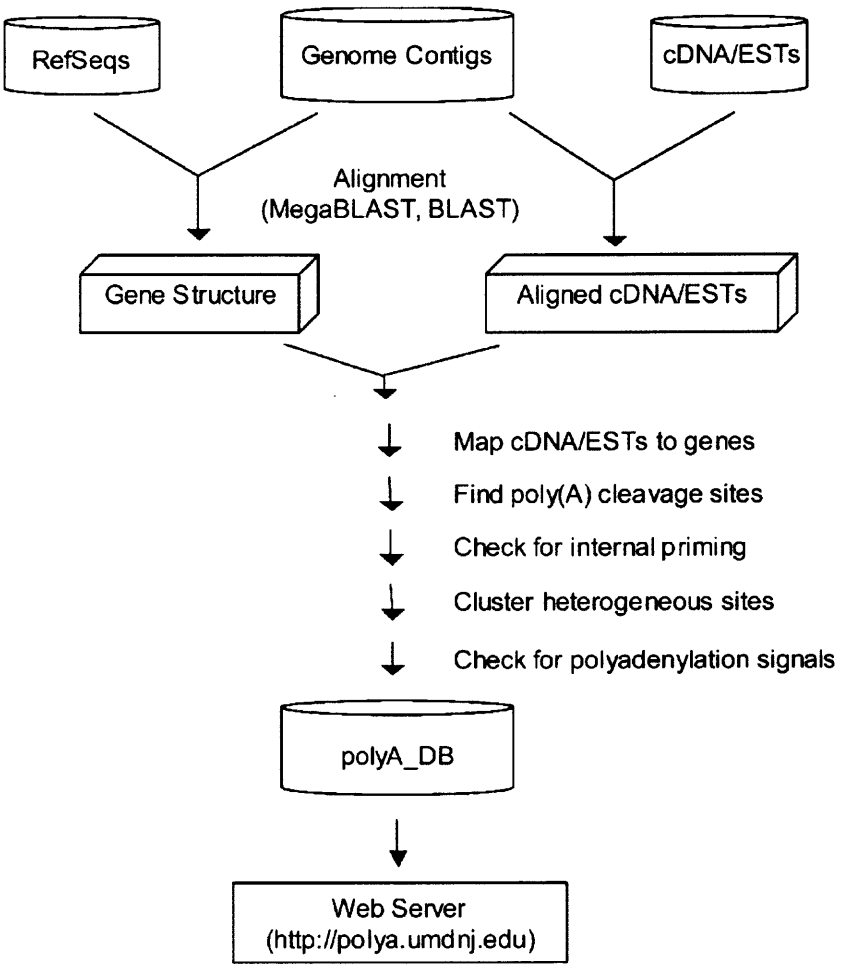


Figure 2.1 An outline of the polyA_DB building pipeline. Arrowed lines indicate the data flow. See text for details.

2.3 Results

2.3.1 Methods and Data Statistics

In the PolyA_DB database, genes are annotated based on LocusLink IDs (Pruitt and Maglott, 2001). A computational pipeline (depicted in Figure 2.1) is designed to accurately map poly(A) sites on genomes:

(1) The genomic location and structure of a gene is determined by the alignment of its RefSeq sequence(s) and genome contigs. In the current version, human genome Build 34.2, mouse genome Build 30, and NCBI March 2004 release of RefSeq mRNAs were used. If a gene has more than one RefSeq sequence, their alignments to genomes are required to overlap, and their transcriptional orientations are required to be the same. The transcriptional orientation of a gene is determined both by its splicing and poly (A) tail information whenever possible. If a gene does not meet these two criteria, it is discarded.

(2) To ensure that only high quality cDNA/EST data are used, the alignment of a cDNA/EST with the genome is required to overlap with that of its corresponding RefSeq. The mapping between cDNA/EST and RefSeq was obtained from the UniGene database (Wheeler et al., 2003).

(3) Only those cDNA/ESTs with poly(A) tails (or poly(T) tails if in anti-sense orientation) are used to infer poly(A) cleavage sites. Poly(A) tails are required to have either 8 or more consecutive A's, or if it has a nucleotide other than A, another 8 or more consecutive A's after that nucleotide is required. Possible internal priming sites are checked by examining the genomic sequence -10 to +10 nt surrounding the

cleavage site. If the sequence has six continuous A's or more than 7 A's in a 10 nt window, it is considered as an internal priming candidate. Poly(A) cleavage sites located within a 24 nt window are considered to be generated from heterogeneous cleavage of mRNA (Pauws et al., 2001), and thus are clustered together.

(4) To further ensure the mapping quality, a genuine polyadenylation site must be either supported by more than one cDNA/EST sequence, or supported by one cDNA/EST alignment together with at least one polyadenylation signal within the upstream –40 to –1 nt region.

Data generated from the pipeline, including genomic locations of poly(A) sites, supporting cDNA/ESTs, number of cleavage sites, and polyadenylation signal information are stored in a relational database using MySQL. Figure 2.2 shows the entity-relationship (ER) schema of the database, which is designed with consideration of the scalability for future expansion of other species. Also in the database are the ortholog gene information obtained from HomoloGene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene), and the tissue/organ information of ESTs derived from cDNA library files from NCBI. Several key data statistics of the database are summarized in Table 2.1.

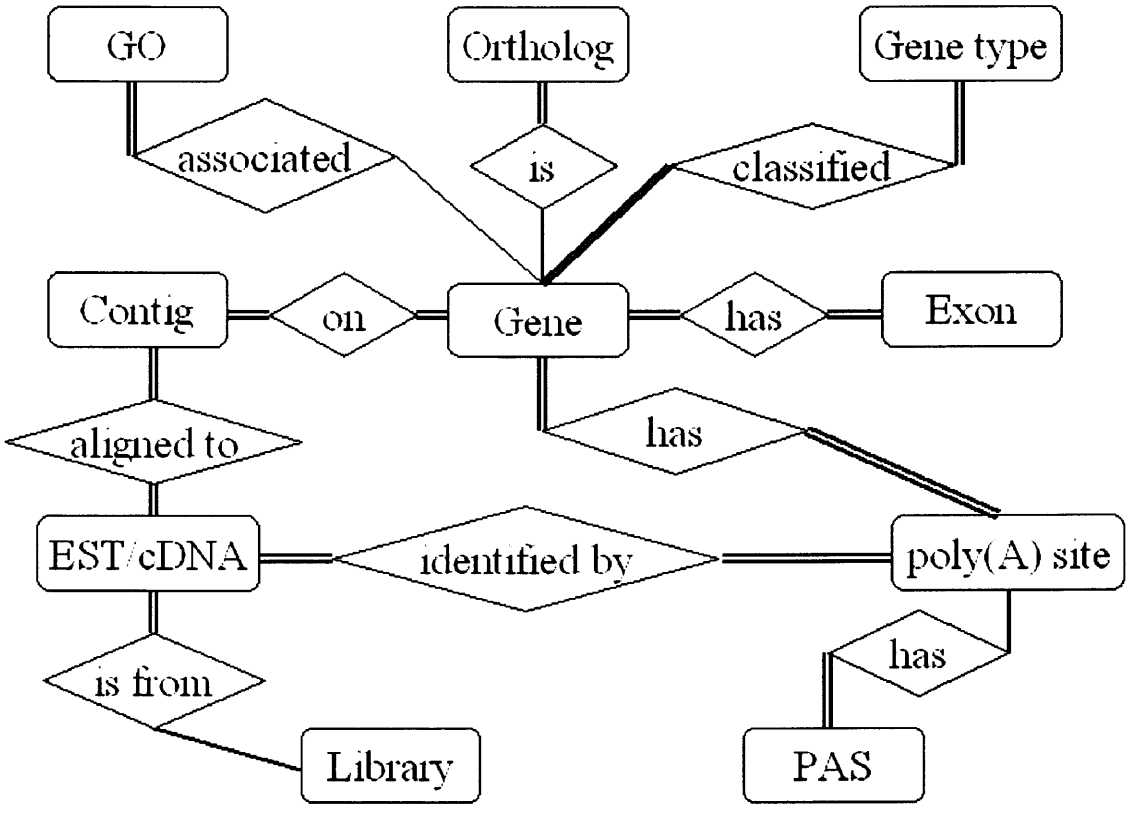


Figure 2.2 Entity-relationship schema of PolyA_DB. Double line stands for full participations, single line stands for partial participations, rounded rectangular stands for database entity, diamond stands for relationship. PAS: polyadenylation signal, EST: expressed sequence tags.

Table 2.1 PolyA_DB Statistics.

	H. sapiens	*M. musculus*	Total
Aligned cDNA/ESTs	2,103,995	1,181,194	3,285,189
Poly(A) sites	29,283	16,282	45,565
genes with one poly(A) site	6,418	7,577	13,995
genes with alternative poly(A) sites	7,524	3,578	11,102
Total Genes	13,942	11,155	25,097
Orthologous pairs	7,935	7,935	7,935
Tissue types[1]	331	155	455
Organ types[1]	107	47	133

[1] It includes diseased tissues and organs. Some tissue and organ types occur in both human and mouse cases.

2.3.2 Data Access and Visualization

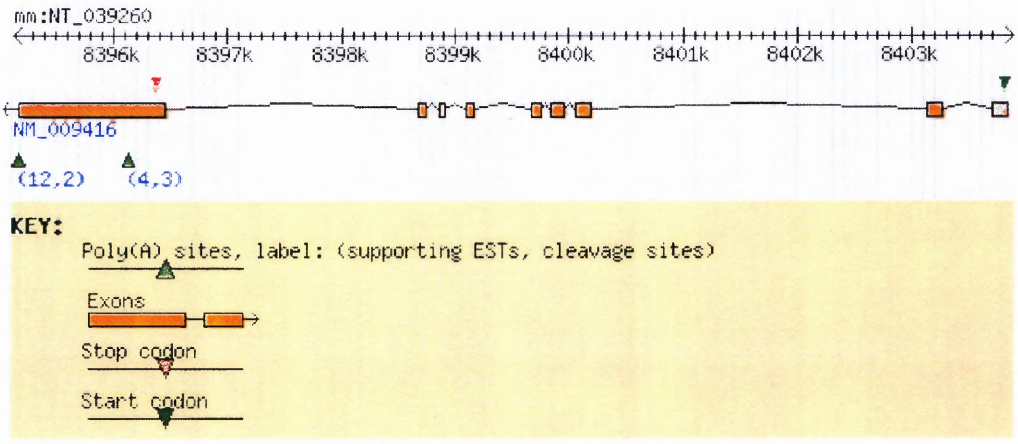
Data and documentation are available from the polyA_DB web server at http://polya.umdnj.edu/polyadb. Data can be either downloaded as MySQL flat file dump or queried through web interface where graphics are dynamically generated using Bioperl modules (Stajich et al., 2002). The web user interface is interactive and provides five basic views:

1. **Gene view.** This view provides a summary of poly(A) site(s) for each queried gene, their positions relative to the RefSeq(s) and the genome contig. The gene structure inferred from the RefSeq(s) and a summary table containing cDNA/EST evidence and number of cleavage sites are also provided. Links for sequence IDs to NCBI resources are provided whenever possible. Figure 2.3(a) shows a polyA_DB gene view of mouse TPM2 gene (b-tropomyosin, LocusLink ID: 22004), with 2 poly(A) sites and their positions, inferred gene structure, start and stop codon positions, number of supporting ESTs, and number of sites generated by heterogeneous cleavage.
2. **Ortholog view.** This view provides a comparison of a pair of human and mouse orthologs. Figure 2.3 (b) shows an ortholog view of mouse TPM2 gene and its human ortholog (LocusLink ID: 7169). The ortholog view readily revealed that the ortholog pair is conserved with respect to both their gene structures and polyadenylation configurations.
3. **Signal view.** This view provides information regarding to *cis*-elements in the surrounding region of a poly(A) site. Currently, PolyA_DB only documents the PAS motif (AAUAAA and its 11 single-nucleotide variants) (Beaudoing et al., 2000) in

the 1-40 nt upstream region of a poly(A) site. Figure 2.3 (c) shows signal views of the mouse and human TPM2 genes, from which the conservation of signal usage of poly(A) sites can be easily revealed.

4. **Evidence view.** This view provides detailed alignment evidence from cDNA/EST sequences, which can be presented by various sorting options including the 3' or 5' position, exon number, cDNA/EST length, and GenBank ID. A table is also provided, which lists all supporting cDNA/ESTs with links to NCBI (Figure 2.4).
5. **Body view.** This view provides tissue/organ information for poly(A) sites (Figure 2.5).

(a)



* Click on the image to see right-click-zoomable SVG image (SVG plugin is needed)

Evidence view | Ortholog View | Signal View | Body View

PolyA_DB report:

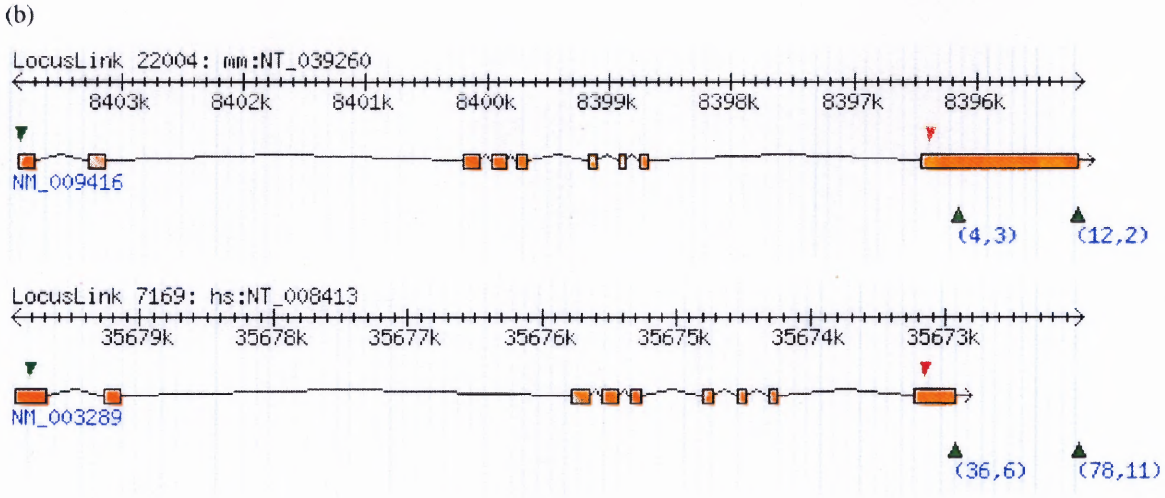
Gene Tpm2 has 2 poly(A) sites. It has a *Homo Sapiens* ortholog. Click here for an ortholog view of polyadenylation.

Organism:	*Mus Musculus*
LocusLink:	22004
Official Symbol:	Tpm2
Gene Name:	tropomyosin 2, beta
RefSeq:	NM_009416
Contig:	NT_039260
UniGene:	Mm.646

Poly(A) sites:

Site ID	Position	# of supporting cDNA/ESTs	# of distinct cleavage site
p.22004.1	8396144	4	3
p.22004.2	8395168	12	2

Figure 2.3 Views offered at the web interface of polyA_DB. (A) Gene view: Mouse gene TPM2 is used as an example. The output includes a pictorial representation of gene structure and poly(A) sites as well as two summary tables regarding the gene and the poly(A) sites. Numbers under each poly(A) site in the picture are the number of supporting cDNA/ESTs (which indicates the relative frequency of use of the poly(A) site) and the number of heterogeneous cleavage sites. (B) Ortholog view of human and mouse TPM2 genes. (C) Signal views of mouse and human TPM2 genes are shown in the upper panel and lower panel, respectively. The position of a signal is relative to the cleavage site, which is set to 0.



(c)

Signal View for LocusLink: 22004

Site ID	Poly(A) signal(s)	Position(s)
p.22004.1	-30 -20 -10 0 AAUAAA	-13
p.22004.2	-30 -20 -10 0 AAUAAA	-15

Signal View for LocusLink: 7169

Site ID	Poly(A) signal(s)	Position(s)
p.7169.1	-30 -20 -10 0 AAUAAA	-12
p.7169.2	-30 -20 -10 0 AAUAAA	-14

Figure 2.4 Views offered at the web interface of polyA_DB. (A) Gene view: Mouse gene TPM2 is used as an example. The output includes a pictorial representation of gene structure and poly(A) sites as well as two summary tables regarding the gene and the poly(A) sites. Numbers under each poly(A) site in the picture are the number of supporting cDNA/ESTs (which indicates the relative frequency of use of the poly(A) site) and the number of heterogeneous cleavage sites. (B) Ortholog view of human and mouse TPM2 genes. (C) Signal views of mouse and human TPM2 genes are shown in the upper panel and lower panel, respectively. The position of a signal is relative to the cleavage site, which is set to 0. (Continued)

Detail View for LocusLink: 22004

poly(A) site ID	Supporting cDNA/ESTs
p.22004.1	W40734, BG793428, BG793519, CB588240
p.22004.2	BY669034, AW048426, BC014809, BC024358, BG793562, BU523778, CD741429, BQ918271, BE993829, CF585091, BM900136, BI104373

*coordinates are relative to contig NT_039260. RefSeq alignments are always shown on top. Identity and e-value statistics are available upon request.

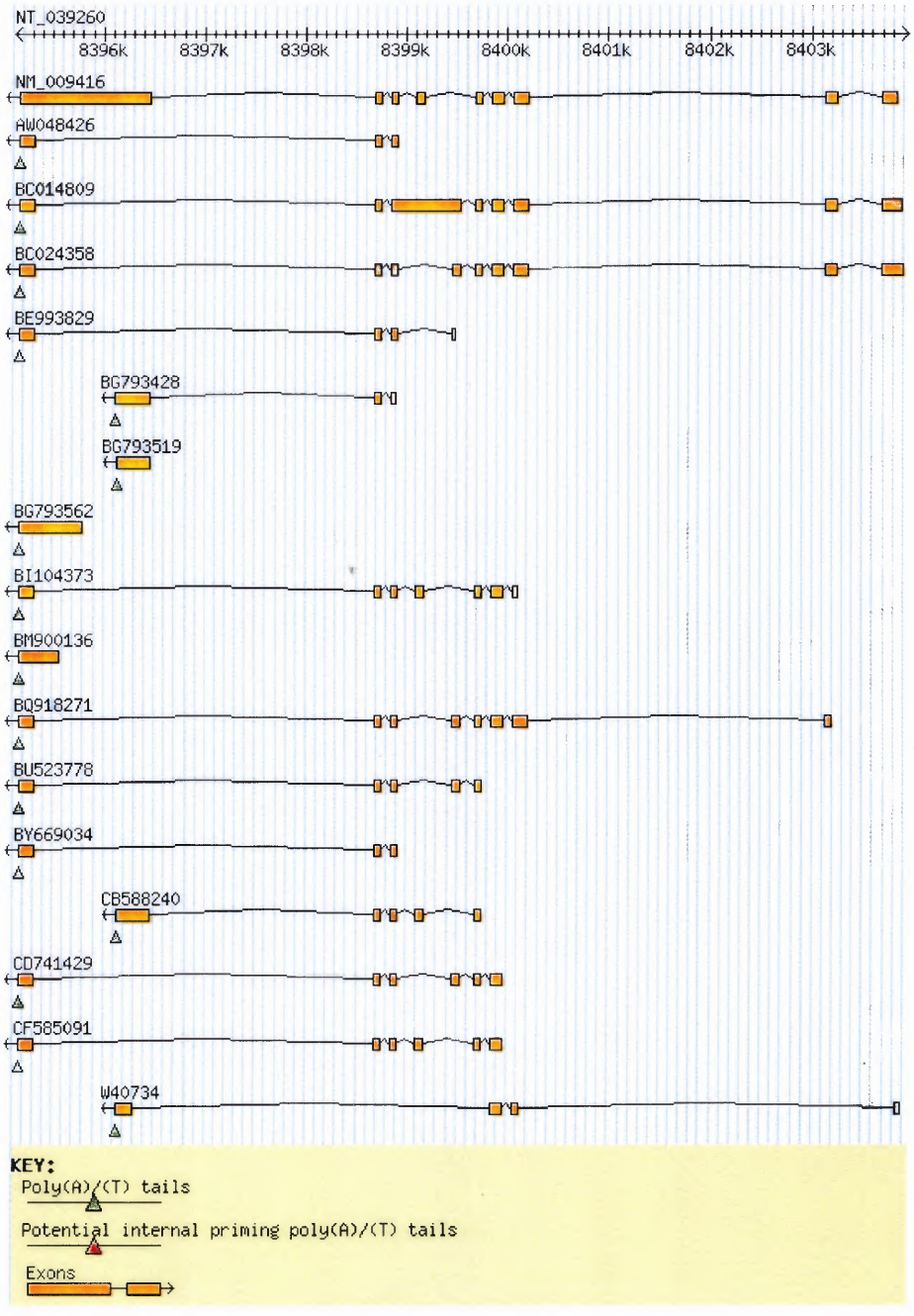


Figure 2.4 Evidence view of polyA_DB. The picture shows detailed alignments for cDNA/ESTs corresponding to mouse TPM2 gene. Only part of the evidence view is displayed, and the alignments are sorted by the number of aligned exons. Only ESTs with poly(A) tails are shown, which are indicated by green arrows.

Body View for LocusLink: 22004

Site ID	Supporting cDNA/EST	Tissue	Organ
p.22004.1	W40734	-	-
p.22004.1	BG793428	Diaphragm/Hind limb muscles	-
p.22004.1	BG793519	Diaphragm/Hind limb muscles	-
p.22004.1	CB588240	embryonic limb, maxilla and mandible	-
p.22004.2	BY669034	Rathke's pouches	-
p.22004.2	AW048426	-	-
p.22004.2	BG793562	Diaphragm/Hind limb muscles	-
p.22004.2	BU523778	-	colon
p.22004.2	CD741429	-	-
p.22004.2	BQ918271	-	colon
p.22004.2	BE993829	-	-
p.22004.2	CF585091	-	pancreas
p.22004.2	BM900136	subfornical organ and postrema	brain
p.22004.2	BI104373	spontaneous tumor, metastatic to mammary. Stem cell origin.	lung

Figure 2.5 Body view of polyA_DB. Body view shows tissue and organ information obtained from ESTs' library information files. The picture shows a body view of mouse TPM2 gene, indicating that both the 5' proximal and 3' proximal poly(A) sites are used in Diaphragm/Hind limb muscles, whereas 3' proximal poly(A) site is used in other additional organs.

2.4 Conclusions

PolyA_DB database, a resource for mammalian mRNA polyadenylation is presented here. This database contains comprehensive information regarding to polyadenylation, including poly(A) sites in the context of gene structure, cDNA/EST evidence for poly(A) sites, polyadenylation signals, conservation of polyadenylation configuration between

orthologs, and tissue/organ information for poly(A) site usage. The information presented in PolyA_DB provides great resources for further large-scale studies to address different questions of polyadenylation systematically. It is a first step towards the systematic view of mammalian mRNA polyadenylation. PolyA_DB will be also of great value to researchers studying the mechanism of polyadenylation and the gene regulation by alternative polyadenylation. PolyA_DB will be continuously updated 1) when new releases of human and mouse genomes and cDNA/EST data are available, and 2) when genome and cDNA/EST data from other species are available for large-scale polyadenylation studies.

CHAPTER 3

CONSERVATION AND GENE ONTOLOGY STUDIES OF POLYADENYLATION IN HUMANS AND MICE

3.1 Abstract

Mammalian mRNA polyadenylation is an important cellular process that affects many aspects of mRNA metabolisms and is tightly controlled by various *cis*-regulatory elements and *trans*-acting protein factors. A great proportion of human and mouse genes have more than one poly(A) site and could undergo alternative polyadenylation. Others have only one poly(A) site. Alternative poly(A) sites could be situated all within the 3' most exon or on different exons of a gene. An interesting question to ask is what is the biological significance of having different polyadenylation configurations such as only one or more than one poly(A) sites. Efforts are made here to address this question by large-scale conservation and gene ontology studies. Firstly, the conservation of three types of polyadenylation configurations between human and mouse ortholog genes are studied. Results show that the conservation between humans and mice is statistically significant, indicating that alternative polyadenylation is a widely employed and conserved cellular mechanism. Secondly, whether these types of polyadenylation configurations are associated with genes involved in certain groups of cellular functions are also examined. Based on associations of Gene Ontology annotations with different polyadenylation configurations, statistical tests identified certain functional groups of genes that are significantly biased towards certain types of polyadenylation configurations. This large-scale study provides important insights into the systematic

view of polyadenylation and the regulation of alternative polyadenylation in mammalian species.

3.2 Introduction

Almost all fully processed mammalian mRNAs have a poly(A) tail at the 3' end, except for most replication-dependent histone genes (Dominski and Marzluff, 1999; Zhao et al., 1999). It is a widely employed cellular process and is interconnected with transcription initiation, splicing, and transcription termination (Calvo and Manley, 2003; Proudfoot, 2004). It has been reported that over 29% of human genes have more than one poly(A) sites and can undergo alternative polyadenylation, resulting in mRNAs with variable ends (Beaudoing et al., 2000). In fact, the number of genes with alternative poly(A) sites documented in PolyA_DB (Chapter 1) are 54% and 32% in humans and mice respectively. The impact of alternative polyadenylation on protein variants has also been studied for a number of genes (Edwalds-Gilbert et al., 1997). It is generally believed that the choice of polyadenylation site is related to tissue types and development stages (Beaudoing and Gautheret, 2001; Edwalds-Gilbert et al., 1997; Zhao et al., 1999).

Alternative poly(A) sites can all situate in the 3'-most exon, giving rise to mRNAs with variable 3' untranslated regions (3'UTRs), or in different exons, leading to mRNAs with variable 3' UTRs as well as distinct protein products (Edwalds-Gilbert et al., 1997). The different types of alternative polyadenylation are referred to as polyadenylation configurations (Figure 1.2). As 3'UTRs often contain sequence motifs that are crucial for mRNA stability, mRNA localization, and protein translation, multiple effects can be implied by the different polyadenylation configurations of genes. A well-

known example of alternative polyadenylation is the IgM heavy chain gene (Takagaki et al., 1996), which has alternative poly(A) sites on different exons. Choice of using one poly(A) site versus another is essential in immune responses by switching from the membrane-bound form to the secreted form.

Because polyadenylation is a widely employed cellular process in mammalian systems, it would be expected that the general mechanism is conserved. In fact, several conservation studies have been done to evaluate the conservation of *cis*-regulatory elements and *trans*-acting factors (Graber et al., 1999b; Legendre and Gautheret, 2003; Proudfoot, 2004). The result of these conservation studies have greatly facilitated the advances in *in silico* predication methods (Graber et al., 1999b; Tabaska and Zhang, 1999). However, whether the polyadenylation configurations of ortholog genes are conserved has not been evaluated. In addition, there has been no attempt to look at whether different polyadenylation configurations are utilized as a general cellular mechanism for certain molecular functions, cellular components, and biological processes. In this chapter, both issues are addressed in humans and mice. 25,097 human and mouse genes with 45,565 polyadenylation sites in PolyA_DB (Chapter 1) are used to study the conservation of polyadenylation configuration between ortholog genes. In addition, the relationships between gene functions (evaluated in three aspects: molecular functions, cellular components, and biological processes) and polyadenylation configurations are addressed. Results show that polyadenylation configurations are significantly conserved between human and mouse ortholog genes; and that polyadenylation configurations are associated with certain functional categories specified by Gene Ontology even after stringent statistical filtering. These comprehensive studies

shed light on the understanding of the regulation of alternative polyadenylation and provide systematic views into the roles of mRNA polyadenylation as a fundamental cellular process in mammalian species.

3.3 Results

3.3.1 Polyadenylation Configurations are Conserved between Humans and Mice

Genes are classified into different types based on their polyadenylation configurations as delineated in Figure 1.2. Briefly, genes with only one poly(A) site are classified as type I genes, genes with multiple poly(A) sites all in the 3'-most exon as type II genes, and genes with multiple poly(A) sites located in different exons as type III genes. Thus type I genes have a single constitutive poly(A) site, whereas type II and III genes have alternative poly(A) sites. 54% of human genes and 32% of mouse genes have multiple poly(A) sites (Table 2.1). To evaluate the conservation of polyadenylation between humans and mice, whether ortholog genes tend to possess the same polyadenylation configurations are evaluated. To this end, genes from the three types of polyadenylation configurations are counted. Counts are then constructed into a contingency table based on the relations of ortholog pair between humans and mice (Table 3.1).

Table 3.1 Conservation of polyadenylation configurations between humans and mice.

Hs \ Mm[1]	Type I[2]	Type II	Type III	Sum
Type I	2,662 (2,193)	494 (865)	177 (275)	3,333
Type II	1,440 (1,747)	1,011 (689)	205 (220)	2,656
Type III	975 (1,137)	497 (448)	256 (143)	1,728
Sum	5,077	2,002	638	7,717
χ^2 test p-value	2.0×10^{-132}			

1. Hs: *Homo sapiens*; Mm: *Mouse musculus*.
2. Values in each cell are in the format of: Observed orthologs (Expected orthologs).

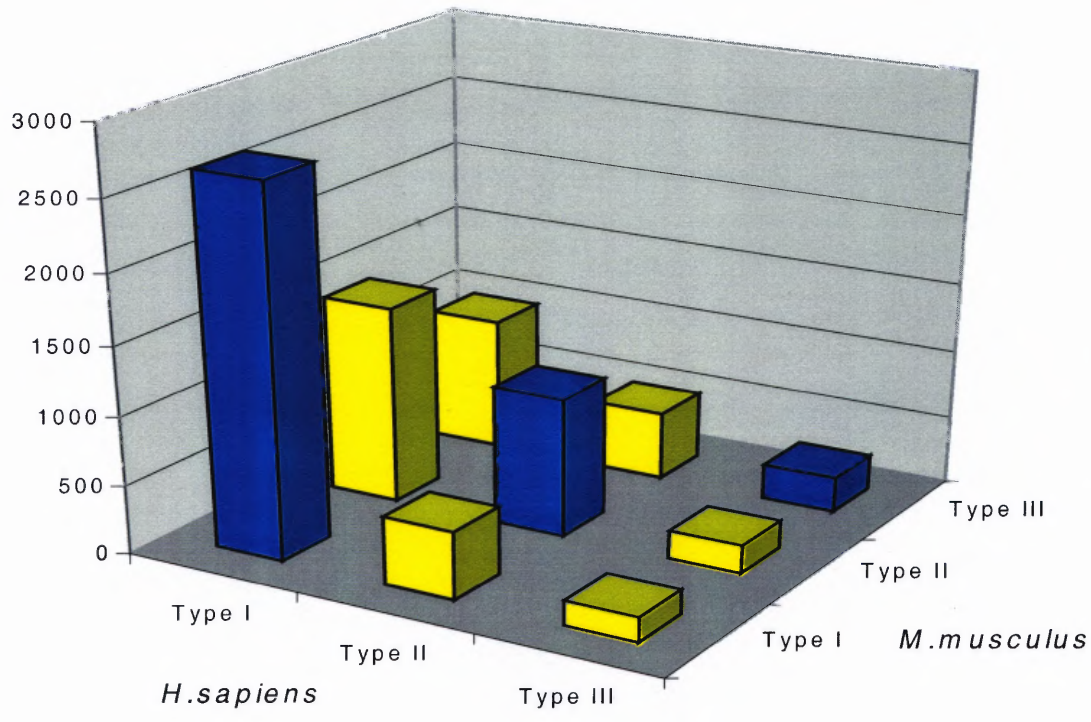


Figure 3.1 Conservation of polyadenylation schemes between humans and mice. Z-axis (vertical axis) is the counts of ortholog pairs. Data is as presented in Table 3.1.

Reciprocal best BLAST hit of protein sequences have identified 7,935 pairs of ortholog genes in humans and mice (see methods). Because some of human and mouse genes don't have enough EST evidence for the mapping of their polyadenylation sites,

only 7,717 out of the 7,935 pair of the ortholog genes are used in the conservation study. The conservation was evaluated with a Pearson χ^2 test applied to the contingency table (Table 3.1). The p-value for the Pearson χ^2 test is 2.0×10^{-132}, suggests that the conservation of polyadenylation configuration between human and mouse orthologs are statistically significant (Figure 3.1). Overall, 51% (3,929 out of 7,717) ortholog gene pairs have conserved polyadenylation configurations. Comparing to expected values computed based on a hypergeometric distribution (see method), number of ortholog genes that are conserved in polyadenylation configurations are all more than 1.2 fold above the corresponding expected values (Figure 3.2). Taken together, these results indicate that alternative polyadenylation is an evolutionarily conserved cellular process between human and mouse ortholog genes.

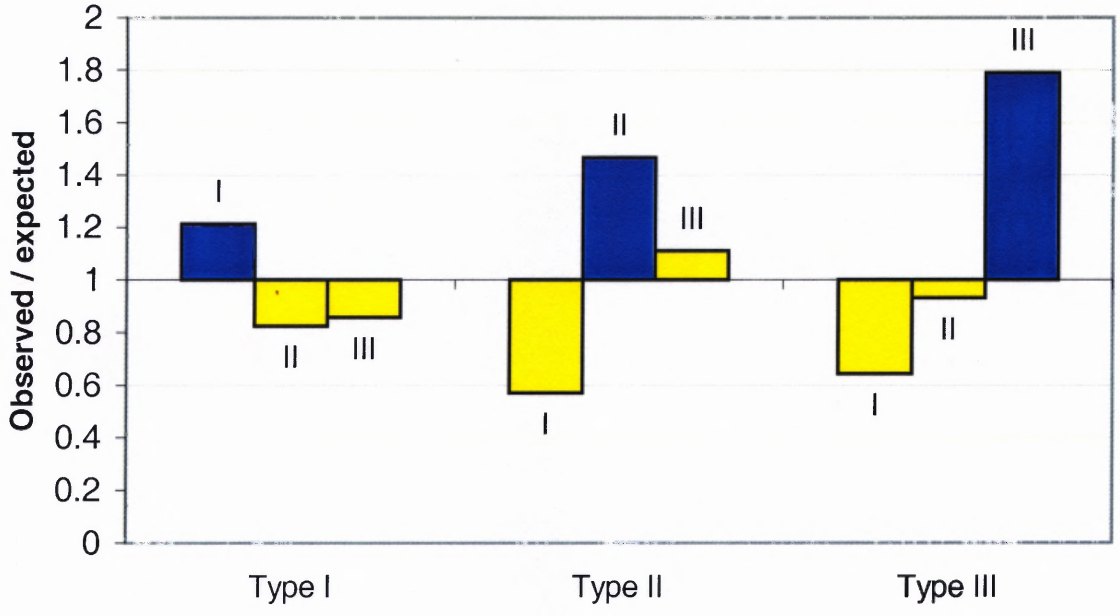


Figure 3.2 Fold changes between observed and expected values of ortholog pairs in humans and mice. Y-axis is fold change computed as observed divided expected values, fold changes greater than 1.2 are colored in blue, others are colored in yellow.

3.3.2 Gene Ontology Study of Polyadenylation Configurations

To address the question of whether polyadenylation configurations are associated with certain functional groups of genes, the relationships between gene functions, as indicated by gene ontology (GO), and the polyadenylation configurations were studied. To identify evolutionarily conserved functional groups, only those ortholog genes that have conserved polyadenylation configurations between humans and mice are used (Figure 3.1). Functional groups of genes are categorized as biological process (BP), cellular component (CC), and molecular function (MF) and are derived from the extensive mapping of LocusLink project at NCBI (Pruitt and Maglott, 2001). Each gene is identified by a LocusLink ID and is annotated in association of several GO terms based on the gene function. Therefore, associations of certain GO terms to polyadenylation configurations are established through the polyadenylation configuration types of associated genes. Each GO term will be voted by the number of genes in each type of polyadenylation configuration. A total number of 7,315 GO terms belonging to the three categories (BP, CC, and MF) were studied. To evaluate whether a GO term is statistically significantly biased towards certain type of polyadenylation configurations, Fisher's Exact test was applied to each GO terms. Stringent false positive rate control was applied using a Benjamanini and Hochberg method (Benjamini and Hochberg, 1995).

In total, there are ten GO terms significantly associated with certain polyadenylation configurations. Genes related to cell surface receptor-linked site transduction (biological process), extracellular genes (cellular component), and site transducer activity (molecular function) are found to be disproportionately associated

with the type I constitutive polyadenylation (Table 3.2). Genes encoding intracellular proteins (cellular component), proteins involved in intracellular protein transport (biological process), and proteins having protein transporter activity (molecular functions) are found to be disproportionately associated with alternative polyadenylation (Type II and Type III). Taken together, it seems that extracellular protein genes tend to have constitutive polyadenylation sites, whereas intracellular protein products are biased to having more than one poly(A) sites, and thus may be more susceptible to the regulation of alternative polyadenylation. In addition, when looking at Type III genes separately, disproportional associations of genes whose protein products are located in the nucleus and have RNA binding activities appear to be statistically significant. Because Type III genes have more than one poly(A) sites located on different exons, this indicates that the regulation of these genes by alternative polyadenylation might be related to splicing.

The results are shown in Table 3.2. Ten GO terms in three categories (Biological Process, Cellular Component and Molecular Function) were shown to have disproportional statistically significant correlations with different types of polyadenylation configurations (Table 1.1). For each GO term, GO ID and annotation are presented on separate lines. A P-value from Fisher's exact test is provided for each GO term in humans and mice, separately. The lower the P-value, the more significant the association is between this GO term and the corresponding tested polyadenylation configuration. GO:0006886 (intracellular protein transport) is associated with both GO:0046907 (intracellular transport) and GO:0015031 (protein transport) through an "is a" relationship (for details see Materials and Methods). Multiple testing adjustment using the Benjamanini and Hochberg method was applied to the selection of significant GO

terms. The numbers of genes associated with a GO term are listed in parentheses in the order of the three types of polyadenylation configurations, i.e. type I, type II, and type III. For example, (171, 37, 3) stands for that 171 type I genes, 37 type II genes and 3 type III genes are annotated as cell surface receptor linked site transduction, identified by GO:0007166.

Table 3.2 Gene Ontology terms disproportionately associated with different types of polyadenylation configuration.

GO terms	Type	*H.sapiens*	*M.musculus*
Biological process			
GO:0007166	I	1.15E–04	9.38E–06
Cell surface receptor linked site transduction		(171, 37, 3)	(154, 30, 1)
GO:0046907	II and III	1.29E–07	1.57E–08
Intracellular transport		(60, 59, 7)	(64, 68, 5)
GO:0015031	II and III	1.41E–07	1.53E–07
Protein transport		(49, 53, 5)	(54, 59, 3)
GO:0006886	II and III	7.25E–08	5.95E–07
Intracellular protein transport		(46, 52, 5)	(51, 55, 3)
Cellular component			
GO:0005576	I	9.98E–05	2.54E–10
Extracellular		(168, 32, 8)	(518, 115, 24)
GO:0005622	II and III	6.32E–08	1.62E–04
Intracellular		(819, 337, 109)	(826, 335, 92)
GO:0005524	III	7.84E–05	2.50E–04
Nucleus		(393, 162, 66)	(376, 143, 53)
Molecular function			
GO:0004871	I	4.94E–06	3.81E–06
Site transducer activity		(344, 77, 18)	(315, 71, 13)
GO:0008565	II and III	4.52E–07	3.41E–07
Protein transporter activity		(30, 42, 1)	(25, 40, 0)
GO:0003723	III	1.17E–05	1.12E–04
RNA binding		(46, 32, 19)	(40, 22, 14)

3.4 Materials and Methods

3.4.1 Conservation Study of Human and Mouse Ortholog Genes

Human and mouse ortholog genes were obtained from NCBI HomoloGene database (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/). Only reciprocal best BLAST hits of protein sequences (7,935 ortholog pairs) were used as human and mouse ortholog genes. Pearson's Chi-squared test was used to test the significance of the conservation of polyadenylation configurations between ortholog pairs. R program (http://www.r-project.org/) was used to conduct statistical test and compute expected values by the Chi-squared test function in the R package. The estimation of expected values based hypergeometric distributions are calculated through the Chi-squared test module in the R package.

3.4.2 Gene Ontology Analysis

Gene ontology (GO) was downloaded from Gene Ontology Consortium website (http://www.geneontology.org/). Gene annotations using established GO terms were obtained from the LocusLink database of NCBI (Pruitt and Maglott, 2001). The full list of GO terms are analysed in three categories, namely Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). Because GO is represented in a tree structure (Ashburner et al., 2000), for each GO term, all associated GO terms were found by a recursive method which searches the whole gene ontology tree for related entries through either "is a" or "part of" relationship. A total number of 7,315 GO terms (3,607 BP, 690 CC, and 3,039 MF) were found to be associated with 9,057 human and 7,700 mouse genes. Fisher's Exact test using 2 by 2 table was applied to assess the bias of

polyadenylation configuration for each GO term. The Benjamini and Hochberg method was applied to eliminate false positives generated as a result of multiple testing (Benjamini and Hochberg, 1995). Briefly, consider m GO terms each with a Fisher's Exact test p-value, and a 0.05 as a significant cutoff, following procedures are applied:

1. Sort p-values for each GO terms from smallest to largest, denote the ith smallest p-value by $p(i)$ for each i between 1 and m.
2. Starting from the largest p-value $p(m)$, compare $p(m)$ with $0.05 \times i/m$, continue as long as $p(i) > 0.05 \times i/m$.
3. Let k be the first time when $p(k)$ is less than or equal to $0.05 \times k/m$, and then declare GO terms corresponding the smallest k p-values are significant.

Two sets of test were carried out independently, which are the test of biased association of GO terms with constitutive poly(A) sites (Type I) versus alternative poly(A) sites (Type II and Type III), and the test of biased association of GO terms with poly(A) sites on different exons (Type III) versus others (Type I and Type II). When testing constitutive poly(A) sites, two columns in fisher 2 by 2 tables are constitutive polyadenylation and alternative polyadenylation, and two rows are "having the GO term" and "not having the GO term". The test for type III gene was carried out in a similar manner except that one column is "is a type III gene", and the other is "is not a type III gene".

3.5 Conclusion and Discussion

In summary, conservation and Gene Ontology studies were carried out to evaluate polyadenylation configurations of genes in humans and mice on a system level. Results

show that polyadenylation configurations appear to be highly conserved between human and mouse ortholog genes; and that statistically biased associations of GO terms and polyadenylation configurations suggest that alternative polyadenylation may be utilized as a general cellular regulation scheme.

Although the polyadenylation configurations are shown here to be highly conserved between human and mouse genes, it is expected that when more EST data (especially mouse EST data) become available, the conservation will be even more significant. As shown in PolyA_DB (Chapter 1), 54% of human genes have alternative poly(A) sites, whereas only 32% of mouse genes have alternative poly(A) sites. The difference is mostly due to the limited supporting EST's from mice (Table 2.1). As EST data is accumulating in a fast rate, it will be more desirable to evaluate conservation among more species when more EST data are available (Table 3.3).

Edwalds-Gilbert *et al.* have detailed alternative polyadenylation into three types: 1) tandem poly(A) sites, 2) coupled with composite in/terminal exons, and 3) coupled with skipped exons (Edwalds-Gilbert et al., 1997). Genes with tandem poly(A) sites are the same as Type II gene classification. However, because the last two types of Edwalds-Gilbert *et al.* are actually results from two distinct mechanisms of alternative splicing, we collectively called them type III genes, where alternative polyadenylation maybe related to splicing. Interestingly, GO study found that RNA binding proteins are disproportionately associated with the type III polyadenylation configuration, indicating a self-regulation mechanism of these genes on mRNA level.

It is also desirable to study the association of different PAS hexamers with polyadenylation configurations. However, association of AAUAAA and its 11 single-

base variants identified by large-scale bioinformatics study doesn't show meaningful statistically significant associations. This is likely due to two reasons: 1) the high dominant occurrence of AAUAAA; 2) association being specified by the spatial arrangement of AAUAAA, GU-rich element, USE, and other auxiliary elements. Future work needs to put PAS, DSE, USE, and other auxiliary elements into context to study their associations with different polyadenylation configurations, probably focusing on tissue- or disease specific regulations.

Table 3.3 Top ten species with higher amount of EST data available at NCBI.

Species	Number of ESTs
Homo sapiens (human)	6,053,112
Mus musculus + domesticus (mouse)	4,333,996
Xenopus tropicalis	887,961
Rattus sp. (rat)	691,985
Ciona intestinalis	684,319
Danio rerio (zebrafish)	592,837
Triticum aestivum (wheat)	587,846
Bos taurus (cattle)	575,247
Gallus gallus (chicken)	531,366
Zea mays (maize)	449,319

Data from http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

Alternative transcription initiation, alternative splicing, and alternative polyadenylation are the major mechanisms that can produce different transcripts from the same gene in mammalian systems. It has been estimated that about 18% of mammalian genes have alternative promoters (Landry et al., 2003; Trinklein et al., 2003) and about 35-55% have alternatively spliced forms (Brett et al., 2000; Kan et al., 2001; Modrek et al., 2001). The estimation of alternative polyadenylated genes is about 32-54% in this study, which is comparable to that of alternative spliced genes. The lower number of

genes with alternative promoters could be attribute to the intrinsic bias of ESTs toward the 3' end of transcripts. As a result, only limited amounts of full-length cDNAs are usually used to estimate alternative transcription initiation events.

Both alternative polyadenylation and alternative transcription initiation can be coupled with alternative splicing. To date, there is no genome-wide conservation study of different types of configurations of alternative transcription initiation that are comparable to the conservation study of alternative polyadenylation discussed in this chapter. However, several individual cases have indicated that such conservation exists between humans and rodents (Landry et al., 2003).

The amount of alternative splicing has been indicated to be comparable among seven different eukaryotes, with no significant differences between humans and other animals (Brett et al., 2002). Several conservation studies have been conducted in recent years, mostly focused on conservation between humans and mice. Because alternative splicing could be very complicated given various combinations of different mechanisms including exon skipping, alternative 3' splice sites, alternative 5' splice sites, and intron retention, there has been no established splicing configurations as comparable to the polyadenylation configurations studied here (Sugnet et al., 2004). Instead, conservation of alternative splicing between human and mouse genomes is often studied by examining alternatively spliced cassette exons (Resch et al., 2004; Yeo et al., 2005) or by comparing sequence features at splice junctions (Kan et al., 2001; Thanaraj et al., 2003). The conservation results estimated from these studies are generally in agreement with each other. Kan et al and Yeo et al estimate that about 10-11% of alternative splicing events are conserved between humans and mice. Thanaraj et al estimates 15% of alternative

spliced junctions are conserved between human and mouse genomes. These estimations are much lower than the conservation of alternative polyadenylation configurations (51%), suggesting that alternative polyadenylation is more conserved than alternative splicing. However, this is somewhat perplexing because alternative splicing events are expected to be under more selective pressure than alternative polyadenylation due to the fact that alternative splicing events are more likely to result in different protein products (Resch et al., 2004). The low conservation of alternative splicing could be a result of limited ESTs that can be used in conservation studies when dealing with multiple species, or could be a result of over-representation of aberrant or species-specific-diseased splicing forms in EST libraries (Sorek et al., 2004). Nonetheless, the high conservation of alternative polyadenylation comparing with alternative splicing does indicate that regulatory sequences on transcripts are very important during evolutionary selections.

CHAPTER 4

TISSUE-SPECIFIC ALTERNATIVE POLYADENYLATION IN HUMANS

4.1 Abstract

Alternative polyadenylation is one of the widespread mechanisms in eukaryotic cells that can be regulated to produce a variety of transcripts from a single gene. More than 50% of genes in humans have multiple polyadenylation sites, the use of which may vary in a tissue-specific fashion. Using a large-scale computational approach, stomach, retina, ovary, and lung show tissue-specific differential usage of strong and weak poly(A) sites. Based on the three types polyadenylation configurations, nervous system, bone marrow, uterus, ear, brain/CNS, pancreatic islet, and ovary are shown to have tissue-level differential usage of poly(A) sites in type II genes (alternate poly(A) sites all located in the 3' most exon); cerebrum, eye, prostate, skin, retina, esophagus, soft tissue, and lung have tissue-level differential usage in type III genes (alternate poly(A) sites locate in different exons); whereas placenta, retina and blood/whole blood show tissue-level differential usage in both types. In addition, positional preferences in these tissue-level differential usage of poly(A) sites are observed. By exploring available microarray expression data, analysis of a set of 20 known polyadenylation-related protein factors identified low concordance of mRNA expression levels in brain tissues compared to other tissues. Examining mRNA levels of these protein factors revealed that PC4 and CstF64-tau message levels are consistently higher in brain tissues, while PTB and U1A messenger levels are consistently lower in brain tissues, compared to other tissues. Finally, enriched *cis*-regulatory elements in poly(A) sites expressed in brain tissues

suggest that interactions between *trans*-acting factors and *cis*-regulatory elements may be very important in tissue-level alternative polyadenylation regulation in brain.

4.2 Introduction

Eukaryotic gene expression could be extremely complicated along the path from a gene to a protein product. At the transcription level, in addition to alternative initiation and alternative splicing, alternative polyadenylation is also frequently used to regulate gene expression by producing mRNA with different 3'UTRs or protein isoforms in a tissue- or stage-specific fashion (Lou et al., 1996; Takagaki et al., 1996). The great contribution of alternative polyadenylation to the complexity of mRNA species underlies the significance of understanding its regulation.

More than 50% of human genes have multiple poly(A) sites, the differential usage of which will result in different protein products or mRNAs with different 3'UTR. Under different physiological conditions, some poly(A) sites might be used more often (strong poly(A) sites) than others (weak poly(A) sites). The 3'-most poly(A) sites are believed to be stronger than those located to the 5' proximal (Beaudoing et al., 2000). It is generally believed that such selection of strong or weak poly(A) sites and/or possible positional preference maybe regulated in a tissue- or disease-dependent fashion (Beaudoing and Gautheret, 2001; Edwalds-Gilbert et al., 1997), and that both tissue-specific *trans*-acting protein factors and corresponding *cis*-regulatory elements are responsible for the regulation of differential selection of poly(A) sites.

The availability of genomic sequences and large amount of cDNA and expressed sequence tags (ESTs) has enabled large-scale study of alternative polyadenylation in the

mammalian transcriptome (Beaudoing et al., 2000; Beaudoing and Gautheret, 2001; Gautheret et al., 1998; Legendre and Gautheret, 2003; Tian et al., 2005; Zhang et al., 2005) as well as other eukaryotic species (Graber et al., 1999a; Graber et al., 1999b). All of these large-scale analyses are based on aligning EST sequences to the genome. As a result, strong and weak poly(A) sites can be identified by the number of supporting ESTs. A set of tissue-specific biases of polyadenylation sites has also been identified (Beaudoing and Gautheret, 2001). However, several issues of tissue-specific alternative polyadenylation events have not been comprehensively addressed, these include but not limited to: 1) Which tissues have biased usage of strong versus weak poly(A) sites? 2) Does tissue-specific alternative polyadenylation have positional preference? 3) To what extent do expression levels of known polyadenylation-related protein factors account for tissue-specific alternative polyadenylation? 4) And whether there are biased *cis*-regulatory elements in the surrounding sequences of tissue-specifically used poly(A) sites?

Efforts were described here of using computational approaches to study tissue-specific alternative polyadenylation events on a large-scale. Tissue-specific biased usage of strong or weak poly(A) sites are observed. The results presented here also suggest positional preferences in the differential usage of poly(A) sites. Analyzing microarray expression data of polyadenylation-related protein factors identified low concordance of expression levels in brain tissues. In addition, polyadenylation-related factors that show consistently higher or consistently lower mRNA levels in brain tissues are identified. Finally, *cis*-elements study in tissue-specifically used poly(A) sites reveals over- and under-represented motifs in specific sequence regions surrounding polyadenylation sites.

Together, the data distinguished several tissues show tissue-specific regulation of alternative polyadenylation and positional effects, and identified candidate *cis*-regulatory elements and *trans*-acting factors that may play important roles in tissue-specific alternative polyadenylation in humans.

4.3 Results and Discussion

4.3.1 Tissue-Specific Usage of Strong and Weak Poly(A) Sites

The processing efficiency of polyadenylation has direct impact on the fate of mRNAs, and eventually affects their protein products. Abnormal processing efficiencies can lead to human diseases such as thrombophilia (Gehring et al., 2001). PolyA_DB has shown that more than 50% of human genes have multiple poly(A) sites and may undergo alternative polyadenylation (Chapter 1). The relative strengths of poly(A) sites of these genes may be different and the usage of alternate sites may be regulated in a tissue- or development-specific fashion (Beaudoing and Gautheret, 2001; Edwalds-Gilbert et al., 1997). For example, regulation of IgM heavy chain secreted form and membrane form during B cell differentiation is achieved by switching the usage between the upstream weak poly(A) site and the downstream strong poly(A) site (Takagaki and Manley, 1998).

With the availability of a large amount of EST data, the polyadenylation processing efficiency can be assessed by the number of ESTs supporting the usage of the corresponding poly(A) sites (Legendre and Gautheret, 2003). The first goal is to identify tissues that show significantly biased usage of strong or weak poly(A) sites. To this end, 22,865 poly(A) sites from 7,524 alternatively polyadenylated human genes in PolyA_DB (Chapter 1) are classified into strong and weak sites by comparing their supporting EST

counts. In consistency with the strategies employed in constructing of PolyA_DB, supporting ESTs are those poly(A)/poly(T)-tailed ESTs after cleaning out possible internal priming candidates. For each gene, a poly(A) site with more than 75% of supporting ESTs is classified as a strong site (3,711 sites), while poly(A) sites with less than 25% of supporting ESTs are classified as weak sites (5,663 sites). It is important to point out that since some cDNA libraries are normalized, EST counts in these libraries cannot be used to reflect polyadenylation efficiency for genes. Therefore, only EST counts from non-normalized libraries are used to infer polyadenylation efficiencies (See method for classifying normalized and non-normalized cDNA libraries). In addition, since this type of analysis relies on a large number of observations to ensure the soundness of statistics, only genes with at least four poly(A)/poly(T)-tailed supporting ESTs were used in this classification step.

The designated tissue types of ESTs were obtained from the extensive annotation of Yeo *et al* (Yeo et al., 2004), based on which, 5,929 out of 9,170 cDNA libraries in PolyA_DB (Chapter 1) can be classified into 48 tissue types. We assess the usage of strong and weak sites by allowing genes in each tissue type to vote for the usage of poly(A) sites based on the number of supporting ESTs. To account for the difference of gene expression levels so that votes are not biased to highly expressed genes, votes for the usage of strong or weak sites are normalized by dividing total number of supporting ESTs for each gene (see method). The tissue-level usage of strong or weak poly(A) sites in each tissue can therefore be measured by summing up votes of all genes studied in the tissue. Based on our classification criteria of strong and weak sites, most tissues are expected to have a linear correlation between votes for strong poly(A) sites versus weak

poly(A) sites (Figure 4.1, R-squared = 0.9758). However, tissues that have differential usage of strong and weak poly(A) sites will show poor correlations. As shown in Figure 4.1, stomach, retina, ovary, and lung clearly deviate from the 95% confidence intervals (orange lines), with stomach, ovary, and lung biased to the usage of strong poly(A) sites and retina biased toward the usage of weak poly(A) sites, suggesting a tissue-level alternative usage of poly(A) sites.

However, pineal gland, muscle, prostate, and eye have only shown marginal deviations (Figure 4.1, black spots). To evaluate the statistical significance of the observed tissue-level biased usage of strong and weak poly(A) sites, Chi-squared tests are further performed on all tissues (see method). To account for the different number of supporting ESTs across human tissues, votes of usage of strong and weak sites are normalized by the total number of studied ESTs in each of the corresponding tissue before statistical tests. As expected, stomach (p=0.034), retina (p=0.022), ovary (p=0.019), and lung (p=0.004) all give significant p-values (Figure 4.1, in brackets), further supporting a tissue-level bias of strong and weak poly(A) sites usage. However, Pearson χ^2 test p-values of marginal tissues are not significant.

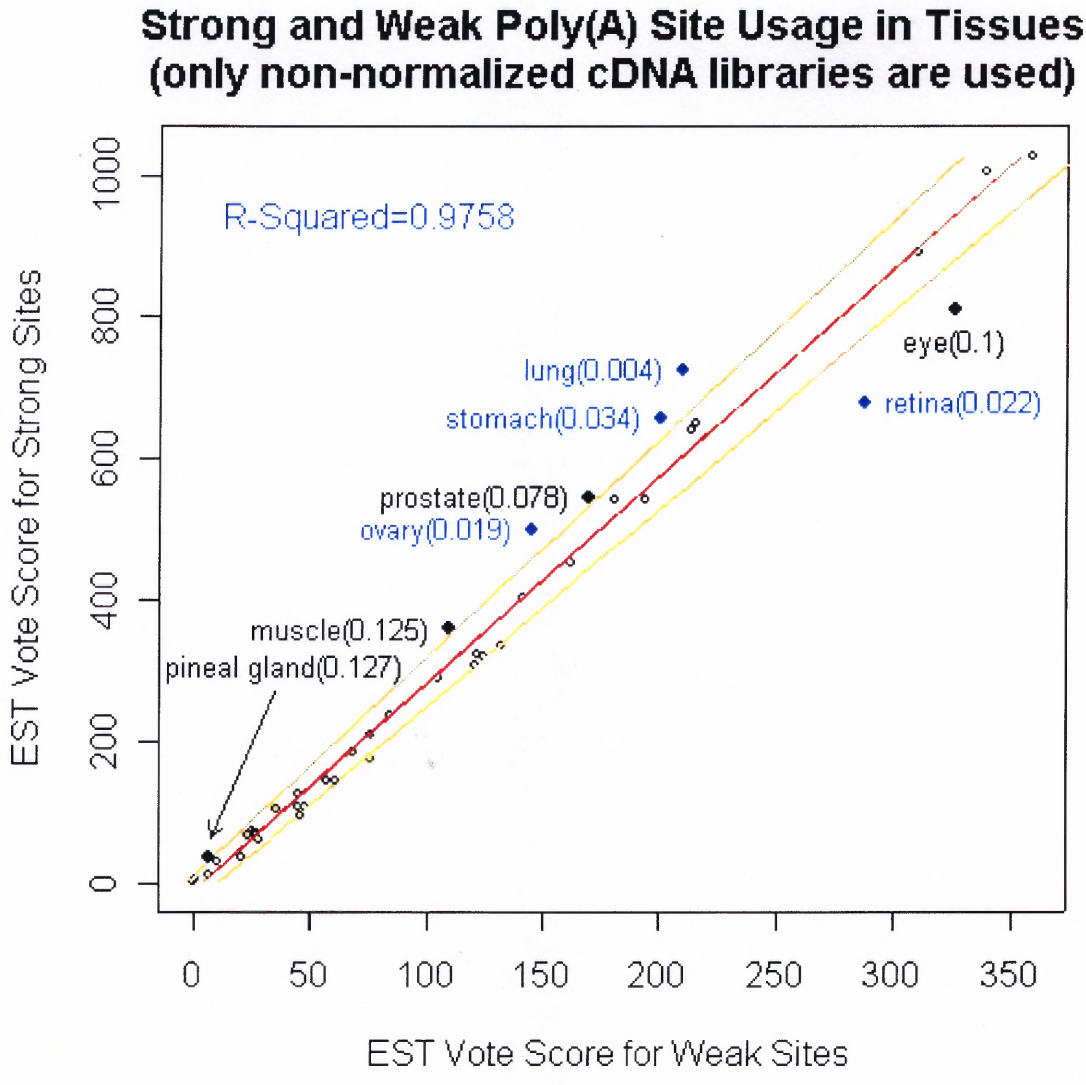


Figure 4.1 Tissue-specific strong and weak poly(A) site usage. Each filled spot or open circle is a tissue type identified by coordinates of vote score for strong sites and vote score for weak sites. Red line is a linear fit with R-Squared = 0.9758. Filled blue spots are tissues with significant Pearson $\chi 2$ test p-values ($p <= 0.05$). Filled black spots are marginal tissues without significant $\chi 2$ test p-values. Filled spots are labelled by tissue names and p-values in brackets.

4.3.2 Positional Preference of Tissue-Specific Alternative Polyadenylation

To address whether there are positional preferences for alternative polyadenylation on a tissue-level, tissue-specific alternative poly(A) site usage is evaluated based on the

relative positions of poly(A) sites on transcripts. To this end, genes are first classified into three types based on their polyadenylation configurations (Chapter 1). Briefly, genes with only one poly(A) site are classified as type I genes, genes with multiple poly(A) sites all in the 3'-most exon as type II genes, and genes with multiple poly(A) sites located in different exons as type III genes. To further study if there is a positional preference of tissue-specific usage of alternate poly(A) sites, poly(A) sites in type II genes are classified into 2F (the 5'-most poly(A) site), 2L (the 3'-most poly(A) site), and 2M (middle poly(A) sites between 2F and 2L); and poly(A) sites in type III genes are classified into 3U (poly(A) sites located upstream of the 3'-most exon) and 3D (poly(A) sites located in the 3'-most exon).

To look for tissue-specific usage of the 3 types of alternate poly(A) sites in type II genes (2F, 2M, and 2L) and the 2 types of alternate poly(A) sites in type III genes (3U and 3D), a voting strategy similar to that was described above for strong and weak poly(A) sites study was applied to type II and type III genes, except that votes now reflect the usage of sites classified by the relative positions located in or upstream of the 3'-most exon, i.e. 2F, 2M, and 2L for type II genes, and 3U and 3D for type III genes. Pearson χ^2 test are then applied to the usage of different types of sites for each tissue, the expected votes are derived from the median votes of all tissues (see method). Statistical significantly disproportional usage of poly(A) sites are observed in both type II and type III genes (Table 4.1, Table 4.2).

Table 4.1 Positional preference of tissue-specific usage of poly(A) site in type II genes.

TISSUE	Gene	EST	Supporting EST vote			Normalized vote			χ^2 test P value	(Observed – Expected) / Expected		
			2F	2M	2L	2F	2M	2L		2F	2M	2L
Blood/Whole Blood	267	797	120.39	52.46	94.15	0.45	0.20	0.35	8.19E-05	0.35	-0.02	-0.25
Ovary	754	1987	290.11	160.77	303.12	0.38	0.21	0.40	9.98E-04	0.15	0.06	-0.14
Retina	1106	2021	419.34	213.41	473.26	0.38	0.19	0.43	3.52E-03	0.14	-0.04	-0.09
Placenta	1428	3066	511.66	315.95	600.39	0.36	0.22	0.42	1.57E-03	0.07	0.10	-0.10
Uterus	1607	4300	487.28	365.07	754.65	0.30	0.23	0.47	6.67E-03	-0.09	0.13	0.00
Brain/CNS	794	1969	239.44	147.82	406.74	0.30	0.19	0.51	4.12E-02	-0.10	-0.07	0.09
Bone Marrow	554	1407	156.00	136.53	261.47	0.28	0.25	0.47	5.92E-03	-0.16	0.23	0.01
Pancreatic Islet	1003	3765	275.13	216.25	511.61	0.27	0.22	0.51	5.03E-04	-0.18	0.07	0.09
Ear	534	1272	129.33	134.22	270.45	0.24	0.25	0.51	1.90E-05	-0.27	0.25	0.08
Nervous	116	137	23.67	30.33	62.00	0.20	0.26	0.53	1.10E-02	-0.39	0.30	0.14

1. Expected usage: 2F: 0.33; 2M: 0.20; 2L: 0.47.

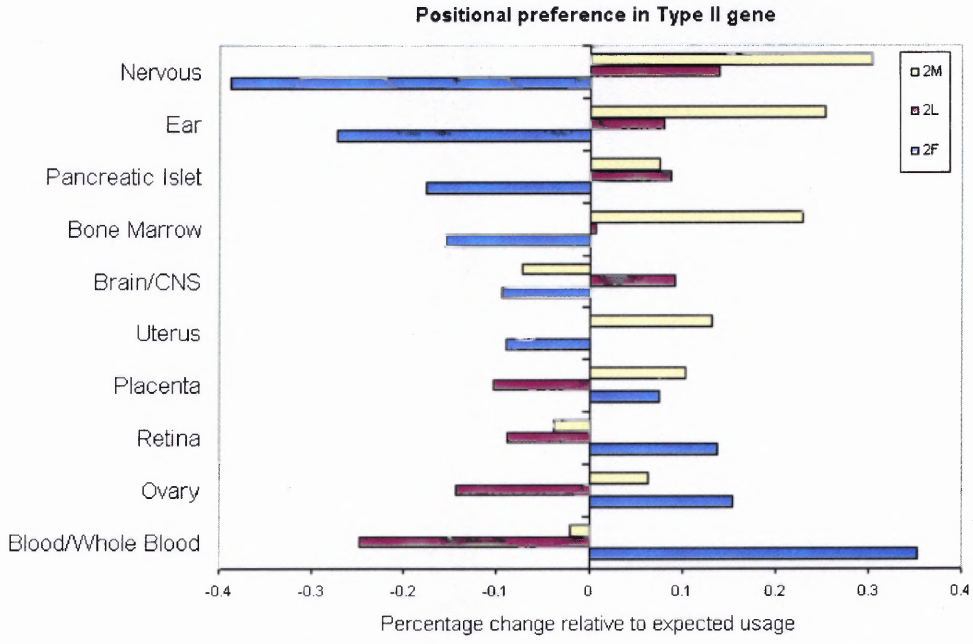
Table 4.2 Positional preference of tissue-specific usage of poly(A) site in type III genes.

TISSUE	Gene	EST	Supporting EST vote		Normalized vote		χ^2 test P value	(Observed – Expected) / Expected	
			3U	3D	3U	3D		3U	3D
Blood/Whole Blood	172	380	37.22	134.78	0.22	0.78	3.09E-05	0.88	-0.11
Retina	777	1391	167.60	609.40	0.22	0.78	1.46E-18	0.87	-0.11
Eye	955	2271	193.63	761.37	0.20	0.80	2.00E-17	0.76	-0.10
Esophagus	73	78	14.00	59.00	0.19	0.81	3.99E-02	0.67	-0.09
Placenta	1008	2164	192.25	815.75	0.19	0.81	5.09E-14	0.66	-0.09
Skin	540	1485	99.00	441.00	0.18	0.82	6.59E-07	0.59	-0.08
Prostate	478	1269	70.76	407.24	0.15	0.85	2.38E-02	0.29	-0.04
Lung	676	1770	97.92	578.08	0.14	0.86	1.52E-02	0.26	-0.03
Soft Tissue	152	202	9.00	143.00	0.06	0.94	3.10E-02	-0.49	0.06
Cerebrum	64	79	2.00	62.00	0.03	0.97	3.56E-02	-0.73	0.09

1. Expected usage: 3U: 0.12; 3D: 0.88.

To further evaluate the positional preferences in observed significant tissues, their votes are compared with expected usage votes computed from median values of votes across all tissue types (Figure 4.2A and 4.2B). Increased usage of 5'-most poly(A) sites (2F) are observed in placenta, retina, blood/whole blood, and ovary, with the decreased usage of 3'-most poly(A) sites (2L), suggesting a shift of usage from 3' to 5' poly(A) sites. However, in bone marrow, uterus, ear, brain/CNS, nervous system, and pancreatic islet the preference is opposite, with an decreased of usage of 5'-most poly(A) sites (2F) and increased usages toward the 3' proximal poly(A) sites (2M and 2L), suggesting s shift of usage from 5' to 3' poly(A) sites. Similarly, placenta, eye, prostate, skin, esophagus, retina, blood/whole blood, and lung show strong increase in the usage of poly(A) sites located upstream of the 3'-most exon (3U), whereas cerebrum and soft tissue show strong decrease of these poly(A) sites but increase usage of poly(A) sites toward the 3' proximal of the mRNA (3D). It is also readily discernable that placenta, retina, and blood/whole blood show positional preferences of usage of poly(A) sites in both type II and type III genes, and more interestingly, the preferences are all toward the 5' most poly(A) sites (2F and 3U). Together, these observations suggest that there are tissue-level positional preferences for the usage of poly(A) sites, such preferences can affect only type II genes, only type III genes, or both type II and type III genes. Since these tissue-specific preferences are observed on a global rather than gene-specific level, the regulation of these preferences may be achieved by coordinating between a set of non-gene-specific *cis*-regulatory elements and a set of *trans*-acting protein factors, the expression level of which might be regulated in a tissue-specific manner.

A.



B.

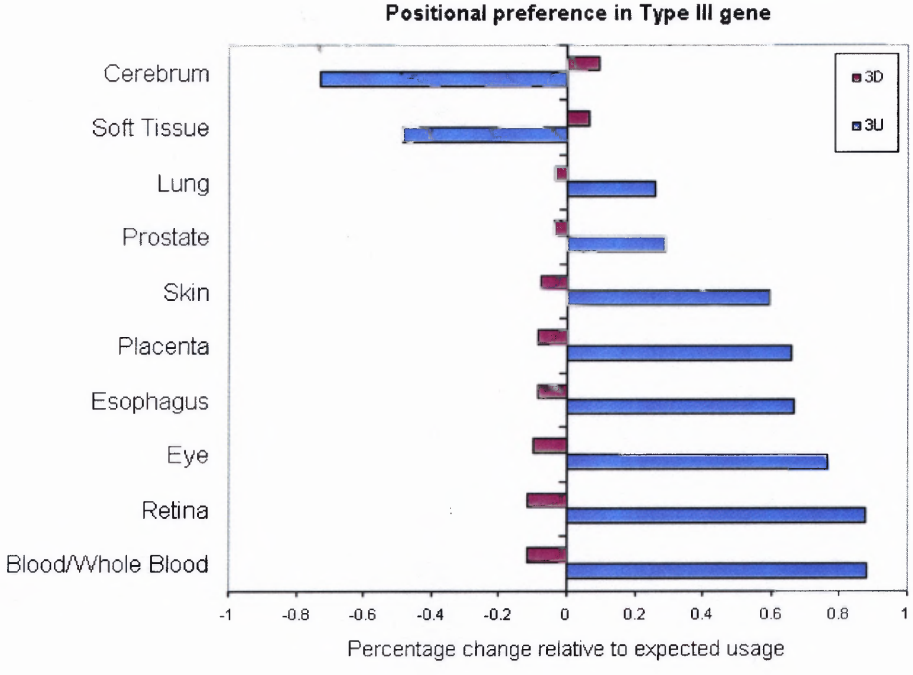


Figure 4.2 Tissue-specific positional preferences of poly(A) site usage. Plotted in both panels are percentage changes relative to expected usage: (observed – expected) / expected. A. Tissue-specific positional preferences of type II genes. B. Tissue-specific positional preferences of type III genes.

4.3.3 Differential Expression of Polyadenylation Related Protein Factors among Tissues

Inspired by Yeo *et al* (Yeo et al., 2004), to further address if there are tissue-specific expression of *trans*-acting factors that might regulate the observed tissue-specific alternative poly(A) site usage, available microarray expression data was analyzed to explore the differences of expression levels of a set of 24 polyadenylation-related factors (Table 4.3) across different tissue types. To date, there are about 41 genes coding for proteins showing evidence of involvement in the polyadenylation process (for the full list Table 1.1), of which this set of 24 factors are known or have been suggested to have regulatory roles in nuclear polyadenylation (for references, see Table 4.3).

Available mRNA expression data were obtained from two independent microarray studies (Su et al., 2002; Su et al., 2004). One set of data is from the chip type HG-U95A, and the other set of data is from the chip type HG-U133A. The two set of independent data were analyzed separately, the comparison of which can be used to validate each other. Therefore, only a set of 25 tissue types that exist in both datasets were selected in this tissue-specific alternative polyadenylation *trans*-acting factor study. The goal is to find tissues that show low concordance of expression levels of trans-acting polyadenylation factors compared to most other tissues, and compare the results with the set of tissues that show positional preferences. Microarray data are first normalized to 75th percentile within each tissue. Out of the 24 polyadenylation-related factors (Table 4.3), 20 of them were represented on both HG-U95A and HG-U133A chips (see Table 4.3 for supporting probe-sets). Therefore, an expression-level vector of 20 polyadenylation-related factors can be obtained for each human tissue. Pearson product

moment correlation coefficients (r) were computed between all pairs of 25 tissues to evaluate the variations of expression in polyadenylation-related factors. If the relative mRNA expression levels across this set of polyadenylation-related factors have a low concordance between two tissues, a low value of r is expected, whereas if the relative mRNA expression levels across this set of factors are similar, a high value of r is expected. Tissues are then clustered based on r-value obtained from both HG-U95A data and HG-U133A data.

Table 4.3 Polyadenylation related protein factors that may play regulatory roles.

Official Symbol	Protein	Gene ID	HG-U95 Probe-sets	HG-U133 Probe-sets
CPSF1	Cleavage and polyadenylation specificity factor 1, 160kDa	29894	33132_at	201638_s_at 201639_s_at 33132_at
CPSF2	Cleavage and polyadenylation specificity factor 2, 100kDa	53981	NA	NA
CPSF3	Cleavage and polyadenylation specificity factor 3, 73kDa	51692	NA	NA
CPSF4	Cleavage and polyadenylation specificity factor 4, 30kDa	10898	35743_at	206688_s_at
CPSF5	CFIM, cleavage factor Im, 25kDa	11051	39142_at	202697_at 213461_at
CPSF6	CFIM, cleavage factor Im, 68kDa	11052	35757_at	202469_s_at 202470_s_at
CSTF1	Cleavage stimulatory factor subunit 1, 50kDa	1477	32723_at	202190_at 32723_at
CSTF2	Cleavage stimulatory factor subunit 2, 64kDa	1478	40334_at	204459_at
CSTF3	Cleavage stimulatory factor subunit 3, 77kDa	1479	41183_at	203947_at
CSTF2T	Cleavage stimulatory factor subunit 2, 64kDa, tau variant	23283	41248_at	212901_s_at 212905_at
HEAB	ATP/GTP binding protein, component of CFIIAm(de Vries et al., 2000)	10978	33149_at	204370_at
PCF11	Pre-mRNA cleavage complex II protein(de Vries et al., 2000; Licatalosi et al., 2002)	51585	41665_at	203378_at
PABPN1	Poly(A) binding protein, nuclear 1(Gunderson et al., 1994; Gunderson et al., 1997)	8106	39050_at	201544_x_at 201545_s_at 213046_at
SYMPK	Symplekin(Takagaki and Manley, 2000; Xing et al., 2004)	8189	32402_s_at	32402_s_at 202339_at
HNRPF	Heterogeneous nuclear ribonucleoprotein F(Veraldi et al., 2001)	3185	38071_at	201376_s_at
HNRPH1	Heterogeneous nuclear ribonucleoprotein H1 (H)(Zarudnaya et al., 2003)	3187	41292_at	201031_s_at 213470_s_at

				213472_at
HNRPH2	Heterogeneous nuclear ribonucleoprotein H2 (H')(Zarudnaya et al., 2003)	3188	41131_f_at 41132_r_at	201132_at
U2AF2	U2 (RNU2) small nuclear RNA auxiliary factor2, U2AF65(Millevoi et al., 2002)	11338	32556_at	218381_s_at 218382_s_at
SNRPA	U1 small nuclear ribonucleoprotein polypeptide A(Gunderson et al., 1994; Gunderson et al., 1997; Lutz and Alwine, 1994; Lutz et al., 1996)	6626	40842_at	201770_at
PC4	Transcriptional coactivator PC4(Calvo and Manley, 2001; Ge and Roeder, 1994)	10923	36171_at	212857_x_at 214512_s_at 221727_at
LOC286528	Similar to HSPC182 protein, human HomoloGene of yeast Ssu72(He et al., 2003)	286528	NA	NA
SFRS3	SRp20(Lou et al., 1998)	6428	351_f_at 40457_at	202899_s_at 208672_s_at 208673_s_at
PTBP1	Polypyrimidine tract binding protein, also known as hnRNP I(Castelo-Branco et al., 2004)	5725	40593_at	202189_x_at 211270_x_at 211271_x_at 212015_x_at 212016_s_at 216306_x_at

1. All factors are described by review articles (Edwalds-Gilbert et al., 1997; Proudfoot, 1996; Proudfoot et al., 2002; Zhao et al., 1999) if not otherwise noted.

The r-value between pairs of tissues computed is plotted as a heat-map in Figure 4.3, the clustering results of tissues are shown on top of the heat-map. Most tissues show high concordance in polyadenylation-related factors expression ($r > 0.75$). However, a distinct cluster (average r-value 0.87 within the cluster) of brain tissues (Amygdala, Thalamus, Caudate Nucleus, Fetal Brain, and Whole Brain) shows a low degree of concordance in the relative mRNA expression levels of polyadenylation-related factors with most other 20 tissues (average r-value 0.55). As shown before in Chapter 4.3.2, brain/CNS and cerebrum show statistical significant positional preference in poly(A) sites usage (Table 4.1, Table 4.2, Figure 4.2). In addition, four other tissues in our microarray data have also shown significant positional effects of poly(A) site usage. These are lung, ovary, placenta, and prostate. As placenta, lung, ovary, and prostate appear to have

positional preference toward the 5' poly(A) sites (2F and 3U), whereas CNS (central nervous system) show preference toward the 3' poly(A) sites, when examining r-values of microarray data between the set of brain tissues with placenta, lung, ovary, and prostate, the correlations between the two groups are found to be very poor, with a mean r-value of 0.41 (Figure 4.3, green blocks). Examining message levels of specific polyadenylation-related factors in the set of brain tissues in both HG-U95A and HG-U133A datasets, relative expression levels of PTB (polypyrimiding tract-binding protein) and U1A (small nuclear ribonucleoprotein polypeptide A) are consistently lower in brain tissues than in other tissues, whereas the relative expression levels of PC4 (transcription coactivator PC4) and CstF64-tau (cleavage stimulatory factor subunit 2 tau variant) are consistently higher in brain tissues (Figure 4.4, Table 4.4). However, comparisons of expression levels between brain tissues and other tissues for other 16 factors don't show such bias in both HG-U95A and HG-U133A datasets (Figure 4.5).

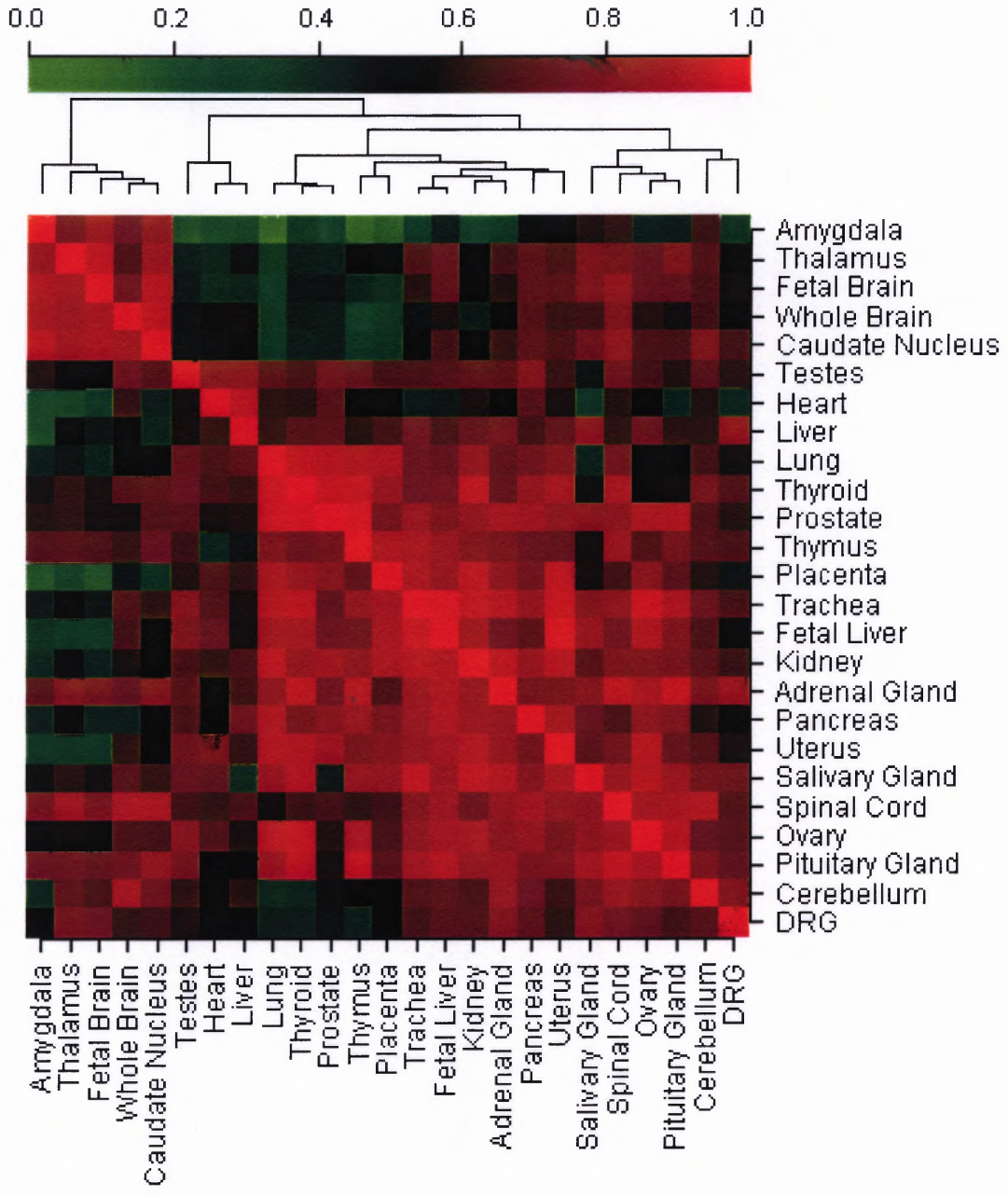


Figure 4.3 Correlation of mRNA expression levels of 20 polyadenylation-related factors (Table 4.1) across 25 human tissues (upper diagonal: data from HG-U133A(Su et al., 2004), lower diagonal: data from HG-U95A(Su et al., 2002), HG-U133A and HG-U95A data are analyzed separately). Based on a scale displayed on top of the figure, small squares are colored to represent the extent of the correlation between mRNA expression patterns of the 20 genes in each pair of human tissues.

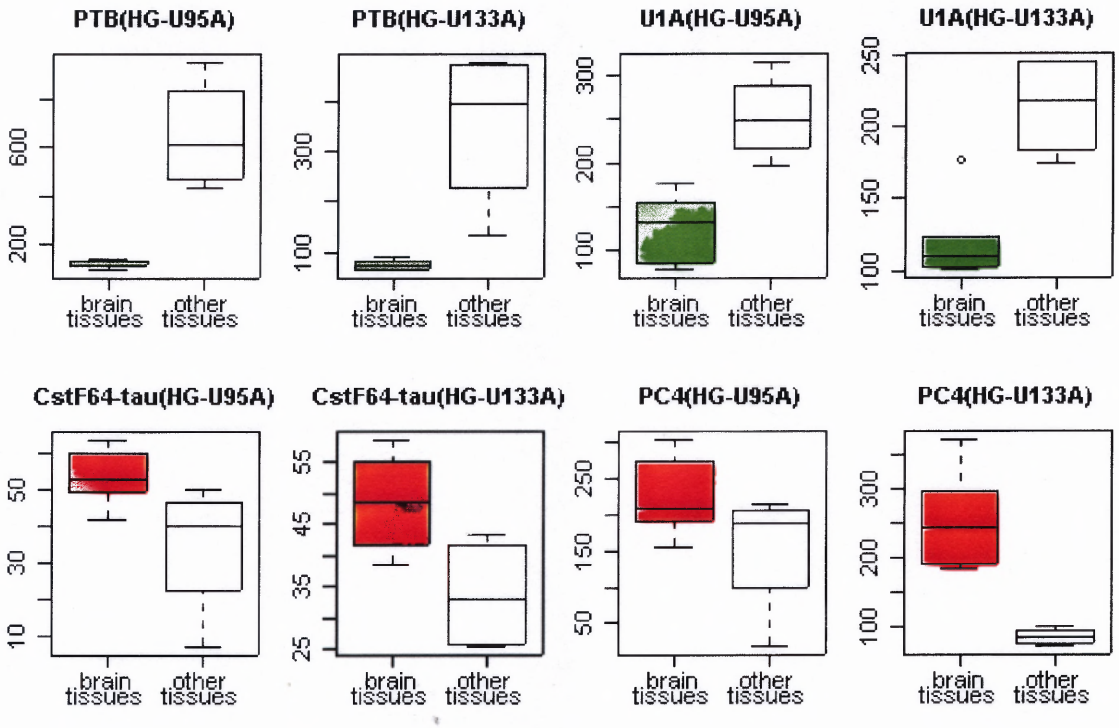


Figure 4.4 Messenger RNA expression levels of PTB, U1A, PC4, CstF64-tau in brain tissues versus in other tissues. Brain tissues include amygdala, thalamus, caudate nucleus, fetal brain, and whole brain. Lower message levels are colored green, whereas higher message levels are colored red. See Table 3 for t-test p values.

Table 4.4 Brain-specific polyadenylation related protein factors.

			PTB	U1A	CstF64-tau	PC4
Rank[1]	HG-U95	Amygdala	1	2	15	22
		Fetal brain	2	7	25	23
		Thalamus	3	1	21	13
		Caudate nucleus	4	3	24	11
		Whole brain	5	9	19	18
	HG-U133	Amygdala	1	3	15	25
		Fetal brain	6	17	17	24
		Thalamus	5	4	21	21
		Caudate nucleus	2	5	24	22
		Whole brain	4	8	22	23
P[2]		HG-U95	2.51×10^{-8}	1.64×10^{-3}	1.04×10^{-3}	6.31×10^{-2}
		HG-U133	6.52×10^{-5}	1.72×10^{-2}	2.15×10^{-2}	8.44×10^{-3}

1. Ascending rank of relative expression levels (25 tissues, maximum rank=25)
2. T-test P values

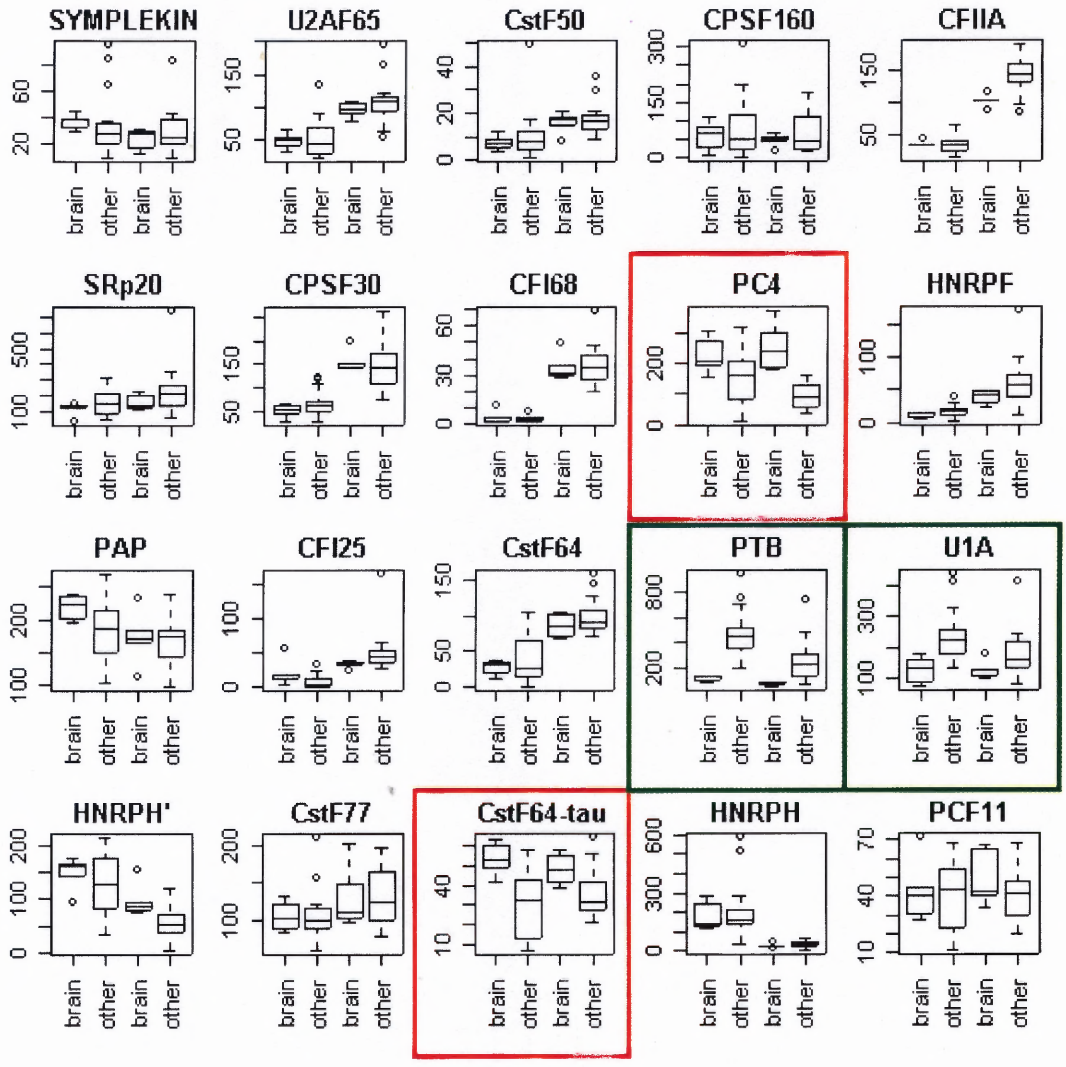


Figure 4.5 Messenger RNA expression levels of all 20 polyadenylation-related factors in brain tissues versus in other tissues. Brain tissues include amygdala, thalamus, caudate nucleus, fetal brain, and whole brain. PTB and U1A have lower message levels and are boxed with green color, whereas CstF64-tau and PC4 have higher message levels and are boxed with red color. From left to right in each boxplot are expression levels from HG-U95A in brain tissues, expression levels from HG-U95A in other tissues, expression levels from HG-U133A in brain tissues, expression levels from HG-U133A in other tissues

PTB is a repressor of mRNA splicing (Chan and Black, 1997; Wollerton et al., 2004) and it has been shown that in brain there are high levels of alternative splicing (especially exon-skippings) (Yeo et al., 2004), which is consistent with our observation

of a low message level of PTB in brain tissues. It has also been shown that PTB can modulate polyadenylation efficiency and this is probably by competing with CstF64 (cleavage stimulatory factor subunit 2, 64KD) for binding to DSE (downstream sequence element) (Castelo-Branco et al., 2004). U1A can increase polyadenylation efficiency by interaction with both USE (upstream sequence element) and CPSF (Lutz and Alwine, 1994; Lutz et al., 1996). An RNA-U1A protein complex containing two U1A molecules can also auto-regulate its production by inhibition of PAP activity (Gunderson et al., 1994; Gunderson et al., 1997). Furthermore, PC4 can regulate polyadenylation by protein-protein interactions with CstF64 (Calvo and Manley, 2001). CstF64-tau is a variant of CstF64 (75% of sequence identity) that is highly expressed in brain and testes (Wallace et al., 1999). Sequence analysis shows that the N-terminus RBD (RNA binding domain) and C-terminus 63 amino acid of putative PC4 binding domain of CstF64 and CstF64-tau are highly conserved (>95% of sequence identity, Figure 4.6), indicating CstF64-tau is also possible to interact with PC4. More interestingly, CstF64-tau gene contains no intron and hence doesn't require mRNA splicing, whereas CstF64 gene has 13 introns and could potentially be regulated at mRNA splicing level, given the unusually high level of alternative splicing in brain tissues (Yeo et al., 2004). Finally, on a similar line of CstF64-tau as a brain-specific counterpart of CstF64, PTB protein also has a brain-specific counterpart nPTB (Ashiya and Grabowski, 1997; Markovtsov et al., 2000). Both nPTB and CstF64-tau message levels are consistently higher in brain tissues compared to other tissues, whereas PTB message levels are consistently lower and there is virtually no difference in CstF64 message level (Figure 4.7). Taken together, this evidence suggests that the observed unusual expression levels of these 4 polyadenylation-

related factors in brain tissues might contribute at least in part to the tissue-level positional effect of poly(A) site usage.

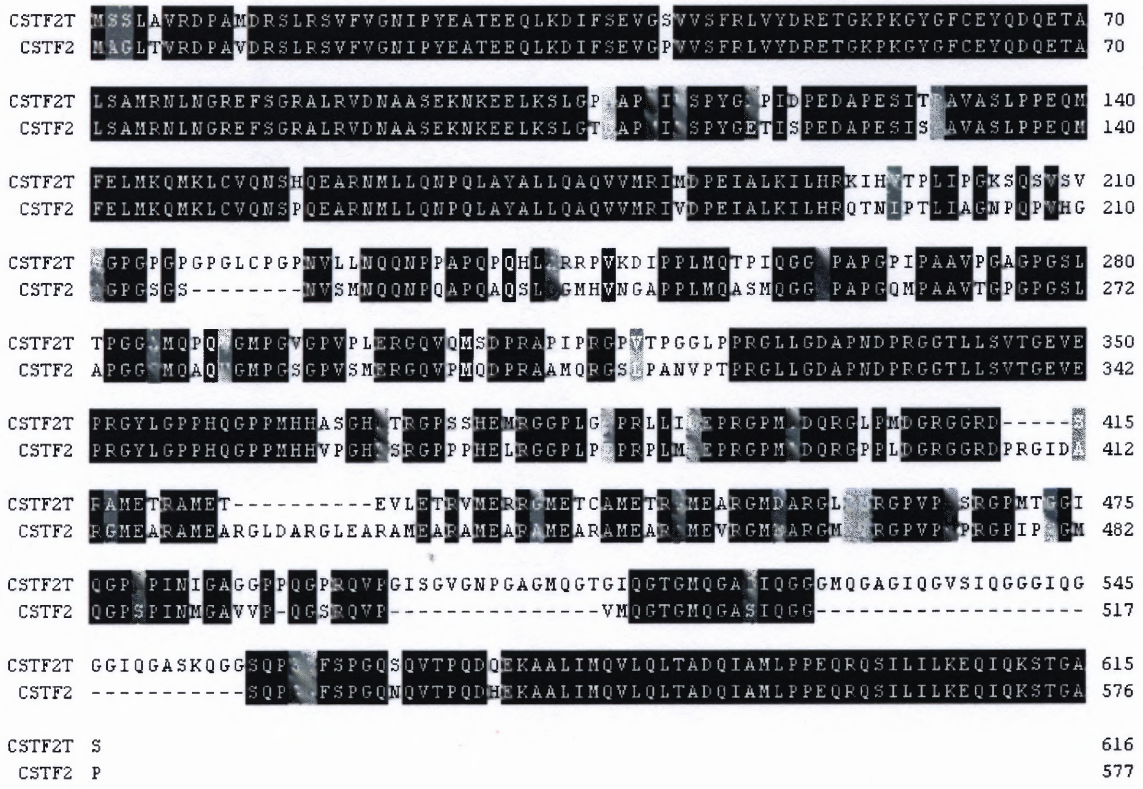


Figure 4.6 Protein sequence alignment of human CstfF64 (CSTF2) and CstF64-tau (CSTF2T).

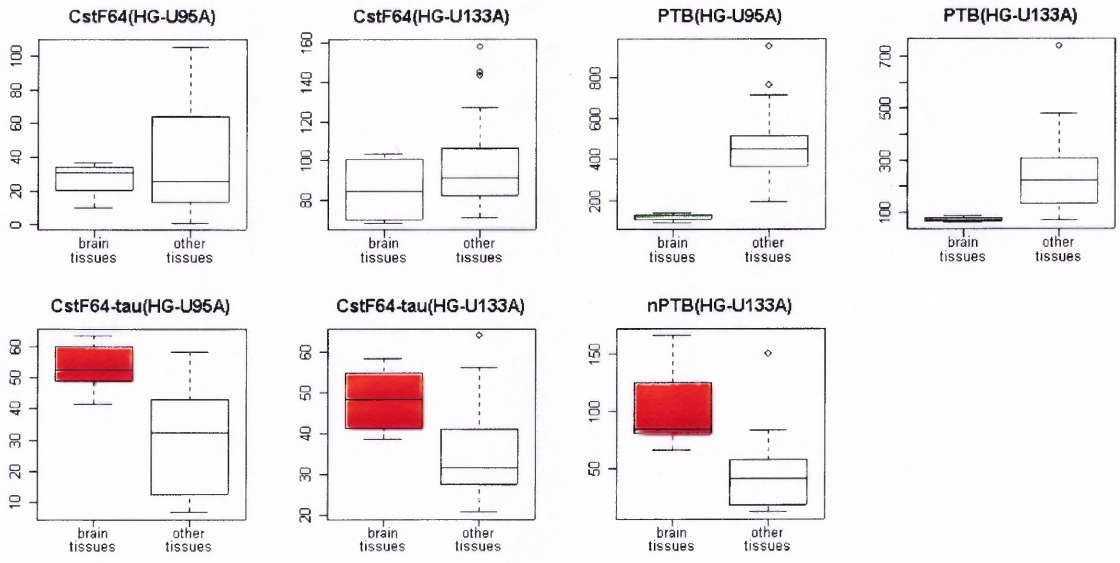


Figure 4.7 Messenger RNA expression levels of CstF64, CstF64-tau, PTB, nPTB in brain tissues versus in other tissues. Brain tissues include amygdala, thalamus, caudate nucleus, fetal brain, and whole brain. Lower message levels are colored green, whereas higher message levels are colored red. Brain-specific nPTB is only represented on HG-U133A chip.

4.3.4 Over- and Under-Represented Motifs in Poly(A) Sites that are Tissue-Specifically Used.

The well-established paradigm in the regulation of alternative polyadenylation is that *trans*-acting factors and *cis*-regulatory elements usually work in concert. The observations of positional effects of alternative polyadenylation and unusual expression levels of several polyadenylation-related factors in brain tissues had prompted a look into candidate *cis*-regulatory elements in brain tissues. Based on known facts of sequence determinants in the poly(A) sites vicinity regions (as discussed in Chapter 1), a poly(A) site vicinity sequence was separated into four regions relative to the cleavage site: upstream 40 nt, where AAUAAA/AUUAAA are usually located, downstream 40 nt, where G/U-rich elements are usually located, 60 nt further upstream where USE maybe

located, and 60 nt further downstream where other auxiliary elements might be located (Figure 6 top, labeled as region "a" to "d" from 5' to 3', also see Figure 1.1). To identify *cis*-regulatory motifs in brain-specific poly(A) sites, an approach was taken that is similar in spirit to the computational method described by Fairbrother *et al.* (Fairbrother et al., 2002). Briefly, all 4,096 (4^6) hexamers were assigned two scores: ΔU, the scaled difference between the frequency of occurrence in the sequence regions (region "a" to "d") from poly(A) sites used in brain and those that are from poly(A) sites not used in brain; and ΔP, the scaled difference between the frequency of occurrence in the vicinity of poly(A) sites and the frequency of occurrence to the distal of poly(A) sites (see method). Therefore one dimension is to identify brain-specific hexamers, and the other dimension is to identify polyadenylation-specific hexamers. The four regions (region "a" to "d") were analyzed separately. Each one of the 4,096 hexamer was then represented on a two-dimensional space identified by coordinates ΔU and ΔP (Figure 4.8). Using a cutoff of 2.5 that corresponds to a p-value of about 0.01 on each dimension, over- and under-represented hexmers were identified. Because the cutoff on each dimension is 0.01, the overall p-value is 10^{-4}, which means number of false positive hexamers identified by this criteria should be less than one ($4{,}096 \times 10^{-4}$). Significant hexamers identified were then clustered based on their nucleotides similarities. Sequence logos (Schneider and Stephens, 1990) were used to build motifs from a multiple sequence alignment of those tightly clustered hexamers (See method).

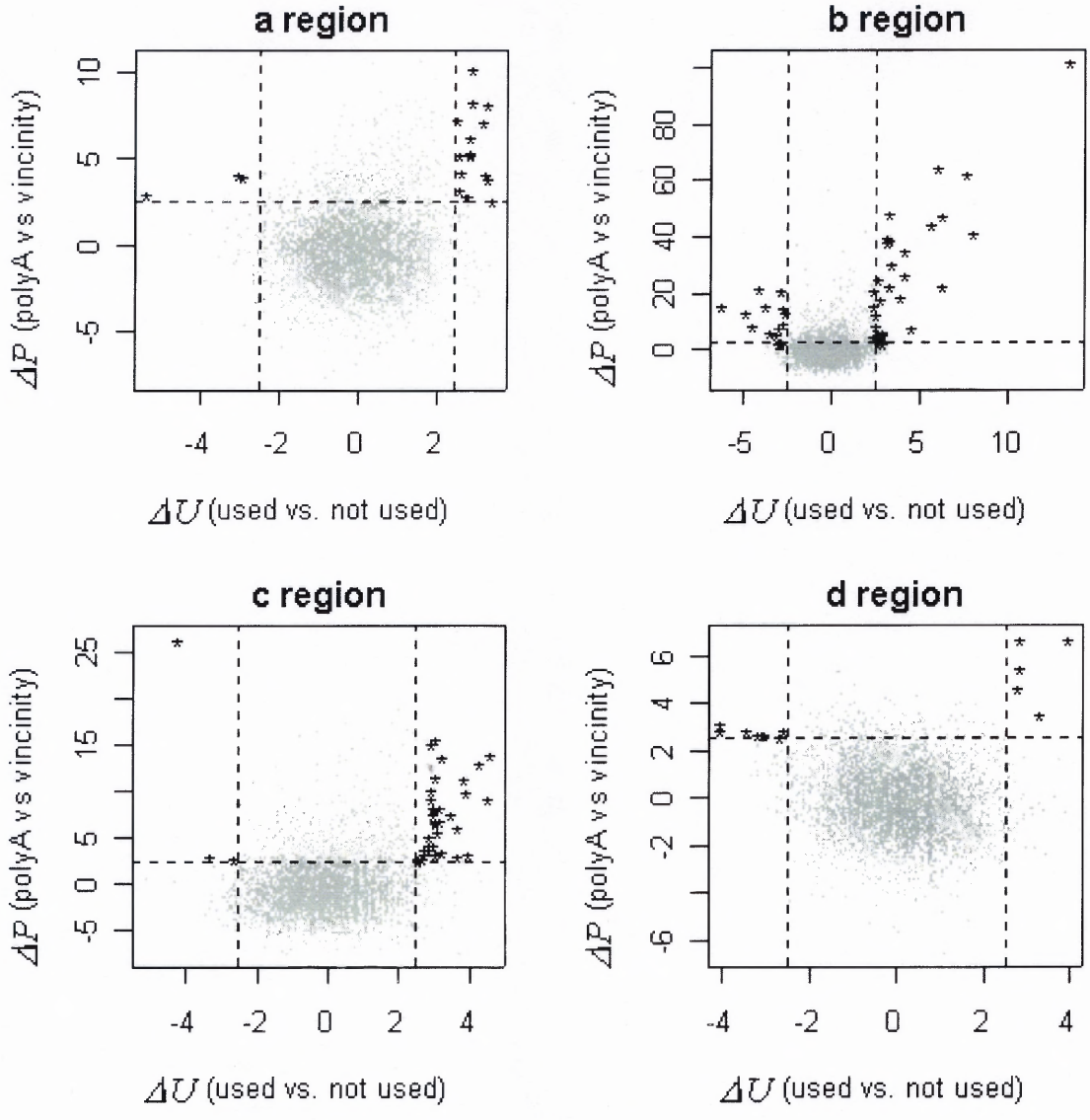


Figure 4.8 Identify brain-specific polyadenylation related hexamers. All hexamers are plotted in the corresponding sub-regions around poly(A) sites. Plot on x-axes are ΔU values, which is comparing poly(A) sites used in brain and those that are not used in brain, on y-axes are ΔP values, which is comparing poly(A) sites vicinity sequences to those distal regions. The cutoffs are both 2.5 as indicated by dashed lines, significant hexamers are depicted as "*", others are depicted as ".".

Nine sequence motifs are identified that are significantly over-represented and three sequence motifs are identified that are under-represented in poly(A) sites that are brain-specifically utilized (Figure 4.9). First of all, UUCU and UUCU are over-

represented in region "a" (USE region) and region "c" (DSE region), which are known to be PTB-binding sites (Castelo-Branco et al., 2004; Garcia-Blanco et al., 1989; Wollerton et al., 2004). In addition, 3 G/U-rich elements were observed in region "c" (UUGUGUGUU, UGGUUGU) and region-"d" (UGUGUGU), which are known downstream elements for enhancing polyadenylation efficiency. Furthermore, an A/U-rich element (AAAUAAAA/UUAA) is over-represented in region-b, which is where polyadenylation signals (AAUAAA or AUUAAA) are located. Very interestingly, the A/U-rich element very much represents a tandemly arranged polyadenylation signal (two copies of AAUAAA), which has been shown before to decrease the usage of upstream poly(A) sites (Denome and Cole, 1988). Finally, a motif of AAAUGU sites was over-represented in region "b" of brain-specific used poly(A), which is know to be an eIF-5α binding site (Xu and Chen, 2001; Xu et al., 2004). Although there is no direct evidence indicating the eIF-5α play any role in polyadenylation, whether these motifs can bind to eIF-5α in vivo might be worth of validating in the future.

When two poly(A) sites are within 100nt distance of each other, AAUAAA signal for the downstream poly(A) sites may lie in region-d of the upstream poly(A) sites. It is therefore reasoned that the under-representation of AAUAAA motif in region "d" is probably due to the shift of poly(A) site usage from upstream to downstream, as being identified in brain and cerebrum (see the positional preference study).

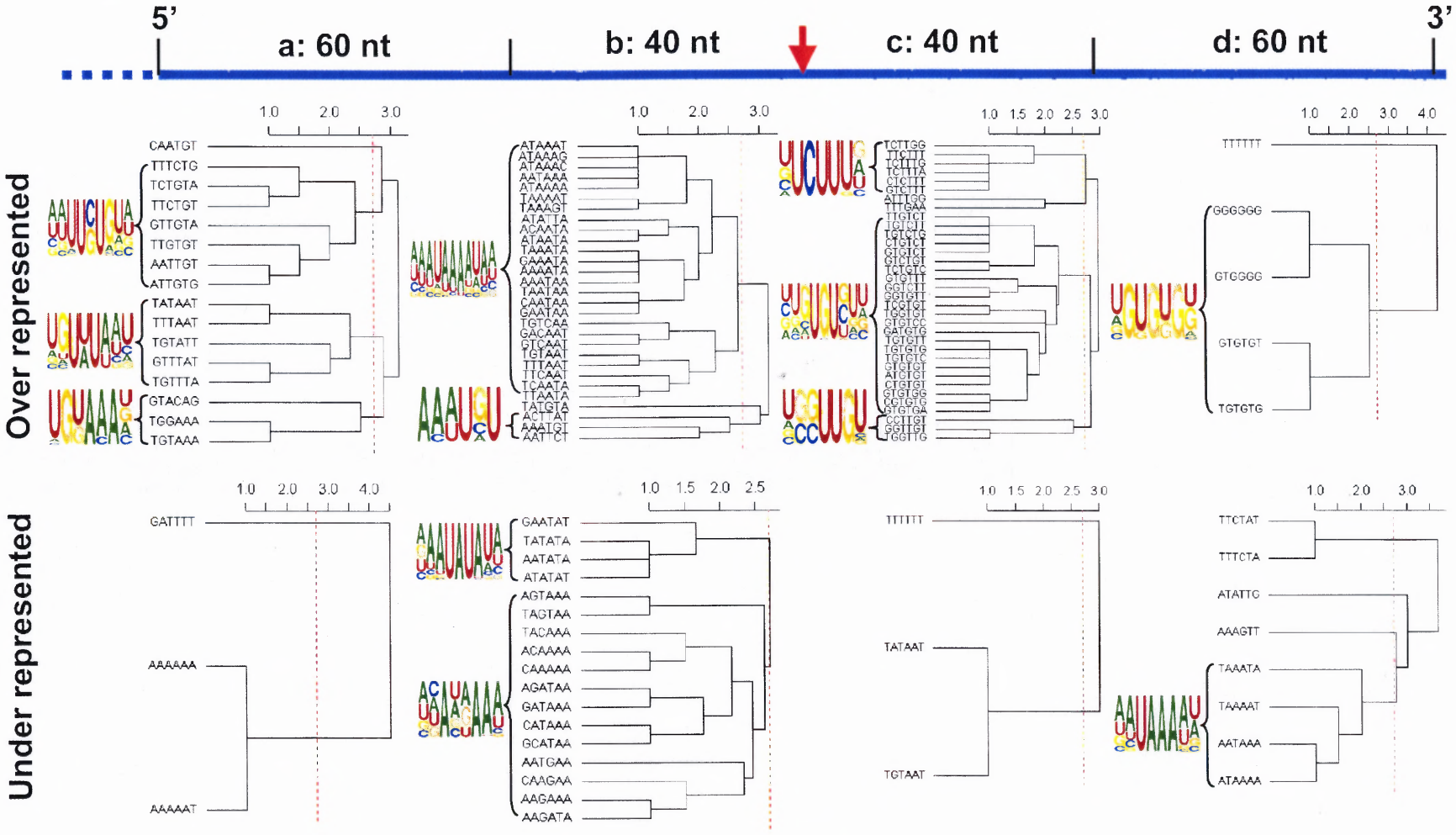


Figure 4.9 Brain specific over- and under-represented motifs. Schematic representations of the 4 sub-regions analyzed are shown in the middle. Brain specific over- and under-represented hexamers are clustered and displayed under each sub-region. The agglomeration method for the hierarchical clustering is "average". Motifs built from each cluster (see method) are displayed as sequence logos.

4.4 Materials and Methods

4.4.1 Data and Resources

Genes with alternative polyadenylation sites, their annotations including poly(A) positions and supporting EST evidence were obtained from PolyA_DB (Chapter 1). Annotations of tissue types for cDNA libraries were downloaded from Yeo *et al.* (http://genes.mit.edu/burgelab/Supplementary/yeo_holste04/Add_Datafile_5). Micro-array datasets (Su et al., 2002; Su et al., 2004) are downloaded from NCBI GEO (Edgar et al., 2002). Mappings of probe-sets to LocusLink IDs are obtained from Affymetrix NetAffix website (http://www.affymetrix.com/Auth/analysis/downloads/taf/HG-U133A_annot_csv.zip, http://www.affymetrix.com/Auth/analysis/downloads/taf/HG_U95Av2_annot_csv.zip). General annotations of cDNA libraries are downloaded from NCBI dbEST FTP site (ftp://ftp.ncbi.nih.gov/repository/dbEST). A PERL script was written to classify cDNA libraries as either normalized or non-normalized. A cDNA library is classified as normalized if its free-text annotation contains "normaliz", but neither "non-normaliz" nor "not normaliz" (All the free text have also been checked for other possible negative phrases such as "without normaliz", "is not/isn't normliz", "are not/aren't normaliz", "has not/hasn't been normaliz", "have not/haven't been normaliz"). Otherwise, the library was considered non-normalized.

4.4.2 Identification of Tissue-Specific Strong and Weak Poly(A) Site Usage

For each gene that has more than one poly(A) site, a poly(A) site is classified as strong if its supporting ESTs are more than 75% of the total number, whereas a poly(A) site is

classified as weak if its supporting ESTs are less than 25%. In PolyA_DB, we only considered poly(A)/(T)-tailed ESTs as supporting evidence for the usage of a poly(A) site. To study biased usage of strong and weak poly(A) site types, each gene will allow to vote for the usage of poly(A) site types according to their supporting EST evidence. In order to avoid the influence of expression level variations, we used an equal-weight voting strategy. The voting scheme is detailed as follows:

$$V_t^p = \frac{1}{G_t}\sum_g V_{g,t}^p \qquad \text{and} \qquad V_{g,t}^p = \frac{E_{g,t}^p}{E_{g,t}}$$

Where:

V_t^p stands for vote for global usage of poly(A) site type p in tissue type t, p $\subset$ (strong, weak);

$V_{g,t}^p$ stands for vote value by gene g for the usage of poly(A) site type p in tissue type t;

G_t stands for total number of genes studied in tissue type t;

$E_{g,t}^p$ stands for number of poly(A)/(T)-tailed ESTs supporting the usage of poly(A) site type p of gene g;

$E_{g,t}$ stands for total number of poly(A)/(T)-tailed ESTs associated with gene g, $E_{g,t} = \sum_p E_{g,t}^p$.

It can be easily derived that $\sum_p V_{g,t}^p = 1$, which means each gene has same weight on voting of a given type of poly(A) site usage so that the vote is not affected by the different expression levels of genes. Expected usage vector was calculated based on the

median vector of all tissue types. Chi-squared test was performed for each tissue with the null hypothesis that the voted vector of poly(A) site usage in the tissue equals the expected usage vector.

4.4.3 Identification of Tissue-Specific Positional Effects

Type II and type III genes are as classified in Chapter 1, Figure 1.2. Poly(A) sites were classified in type II genes into 2F (the 5'-most poly(A) site), 2L (the 3'-most poly(A) site), and 2M (middle poly(A) sites between 2F and 2L); and classified poly(A) sites in type III genes into 3U (poly(A) sites located upstream of the 3'-most exon) and 3D (poly(A) sites located in the 3'-most exon). The exon/intron structures are derived from RefSeq of each gene, as identified by LocusLink IDs (Pruitt and Maglott, 2001). A voting and chi-squared test strategy similar as described above was used to evaluate tissue-specific positional effects of poly(A) site usages. Except in the current analysis, instead of strong or weak poly(A) site types, poly(A) site types are $p \subset$ (2F, 2M, 2L) for type II genes, and $p \subset$ (3U, 3D) for type III genes. Type II and type III genes are analyzed separately.

4.4.4 Microarray Data Analysis of *Trans*-Acting Factors

Microarray data from HG-U95A and HG-U133A were analyzed separately. Expression levels of probes were inferred from the average difference values (AD). AD values are first normalized to 75th percentile within each chip. When more than one probes mapped to the same gene, median value was taken as the gene expression level. For tissue types with more than one sample, median values are taken. Pearson product-moment

correlation coefficients (r-value) between all pairs of tissues were calculated with the vector of expression levels from the 20 genes (Table 4.3). Tissues are clustered using 1-r as distance between a pair of tissues.

4.4.5 Identification of Putative *Cis*-Regulatory Motifs

Previous researches have shown that the -100 to +100 region relative to a poly(A) site has different nucleotide composition than regions further upstream (<-100) or downstream (>+100) of a poly(A) site (Legendre and Gautheret, 2003; Tian et al., 2005). We therefore define the -100 to +100 nt region as a poly(A) vicinity region, the -300 to -200 and +200 to +300 regions as poly(A) distal regions. A poly(A) vicinity region is further separated into four sub-regions: upstream 40 nt, where AAUAAA/AUUAAA are usually located, downstream 40 nt, where G/U-rich elements are usually located, 60 nt further upstream where USE maybe located, and 60 nt further downstream where other auxiliary elements might be located (Figure 4.8 top, labeled as region a to d from 5' to 3'). Frequency of occurrence of all 4,096 hexamers were checked in each sub-region and in poly(A) distal regions as well. Based on these frequency of occurrence, all 4,096 hexamers were assigned two scores in each sub-region: ΔU, the scaled difference between the frequency of occurrence in the sub-region from poly(A) sites used in brain and those that are from poly(A) sites not used in brain; and ΔP, the scaled difference between the frequency of occurrence in the sub-regions and the frequency of occurrence in the distal regions. The calculations of ΔU and ΔP are as those in Fairbrother *et al.* (Fairbrother et al., 2002). Briefly as depicted below: consider a dataset from region "a" of poly(A) sites that are used in brain tissues containing N_U sequences, and a dataset from

region "a" of poly(A) sties that are not used in brain tissues containing N_N sequences. Then for a hexamer with occurrence of O_U in the brain-used dataset, and O_N in the brain-not-used dataset, $\Delta U = \frac{O_U / N_U - O_N / N_N}{\sqrt{(\frac{1}{N_U} + \frac{1}{N_N}) g(1-g)}}$ where $g = \frac{O_U + O_N}{N_U + N_N}$. ΔP is defined similarly, with the set of sequences not used in brain replaced by distal sequences of all poly(A) sites. ΔU and ΔP are then calculated for each sub-region "a" to "d".

Each hexamer was then represented on a two-dimensional space identified by coordinates ΔU and ΔP. Using a cutoff of 2.5 that corresponds to a P-value of about 0.01 on each dimension, over- and under-represented hexmers were identified. Clustering of hexamers, alignments of hexermers, and generation of motifs using pseudocounts to pad edge positions are essentially follow those are described in Fairbrother *et al.* (Fairbrother et al., 2002), except that the cutoff for hexamer clusters for clustalW alignments was that the cluster must have 3 or more members.

CHAPTER 5

A SAGE VIEW OF HUMAN ALTERNATIVE POLYADENYLATION

5.1 Abstract

Alternative polyadenylation is a major post-transcriptional regulatory process that affects a large amount of higher eukaryotic genes. Human SAGE data from 241 libraries was used to search for alternatively polyadenylated genes on a large scale. In total 35,547 human genes were studied, 20,906 genes were shown to have multiple heterogeneous SAGE tags detected in at least one of the 241 libraries. Combining with EST data, we identified 7,196 alternatively polyadenylated genes supported by both EST and SAGE data. Furthermore, SAGE tags were mapped for 31,042 human poly(A) sites based on supporting EST sequences. The selection of different poly(A) sites were demonstrated for two genes, PAPII and CstF-77, with instances of library-specific differential usage of poly(A) sites. This demonstrated that in addition to analyzing gene expression profiles, SAGE data could also be used to study alternative polyadenylation.

5.2 Introduction

More than 29% of human mRNAs have multiple polyadenylation sites (Beaudoing and Gautheret, 2001). Selection of different poly(A) sites could produce transcripts with different 3' ends and may affect the stability and/or translatability of mRNAs (Curtis et al., 1995; Ross, 1995). It has also been shown that differential polyadenylation tends to occur in a disease- or tissue-specific manner (Beaudoing and Gautheret, 2001; Edwalds-Gilbert et al., 1997). A complete survey of alternative polyadenylation in human has been

carried out using Expressed Sequence Tag (EST) data (Gautheret et al., 1998). However, because of known issues of EST data, such as sequencing errors, internal priming, presence of chimeric ESTs and paralog genes, potential vector contamination at the ends, and inclusion of genomic sequence because of incomplete splicing, EST-based alternative polyadenylation survey must employ several limitations to ensure high specificity. Criteria applied currently using EST-based methods are: 1) using poly(A)/(T) tailed ESTs; 2) looking for the existence of putative polyadenylation signals; and 3) looking for poly(A) sites with more than one EST supports. As a result, not all genes can be studied. In addition, because ESTs are developed for gene discoveries rather than for quantitative measurements of transcripts, it is desirable to look for other large-scale expression data for evidence to study alternative polyadenylation, the result of which can then be combined with the current EST-based analysis to refine our systematic view of alternative polyadenylation in humans.

Serial Analysis of Gene Expression (SAGE) is a method that detects and quantifies the expression of large numbers of transcripts (Velculescu et al., 1995). Recently, a great amount of SAGE information has been collected and is made available for public analysis (Lash et al., 2000). Briefly, SAGE involves two major principles: 1) a short sequence (SAGE tag) of 10 base pairs (bp) is sufficient to identify uniquely an expressed transcript given it is obtained from a defined position within the transcript. To this end, an anchoring enzyme (AE, usually is NlaIII, of which the restriction site is CATG) is used to facilitate the anchoring of the 3' most tag prior to the poly (A) tail of the transcript. 2) the abundance of a transcript is directly proportional to the number of times the corresponding SAGE tag is observed (Velculescu et al., 1995). Currently, the

major applications of SAGE are in the fields of novel gene discovery and gene expression profiling from cell populations of different physiological conditions. Because SAGE tags are sampled from the 3' end of transcripts, it is expected to carry a vast amount of information about polyadenylation variants. Presented here is an extensive survey of human alternative polyadenylation combining SAGE and EST data for the first time.

5.3 Results and Discussion

5.3.1 Alternative Polyadenylation Events Result in Heterogeneous SAGE Tags

Usage of different poly(A) sites can result in transcripts with different 3' regions under different physiological conditions. As the anchoring enzyme NlaIII recognizes a 4 nt sequence (CATG), the recognition site is expected to occur on average at least once in every 256 nt (4^4). Considering the base composition in the 3'UTR region computed from UTRdb (Pesole et al 1998), which is A: 28%, T: 28%, G: 22%, C: 22%, the recognition site is expected to occur on average at least once in every 264bp. A survey on the distances between different polyadenylation sites in PolyA_DB shows ~80% of them are greater than 256 nt, with an average distance of ~1300 nt, of which the probability of observing at least one CATG is $1 - (1 - 4^{-4})^{1300-3} = 0.9938$ (0.9928 when considering base composition in the 3'UTR region). It is thus highly likely to obtain different SAGE tags from different forms of transcripts resulted from alternative polyadenylation. In fact, direct search of NlaIII recognition sites in the inter-poly(A) sites regions (sequence regions between all pair of alternative poly(A) sites) of 7,524 human alternative polyadenylated genes in PolyA_DB (Chapter 1) demonstrated that CATG occurs 78.92%

(CATG in 21,219 out of 26,888 inter-poly(A) sites sequences). Therefore, most of alternative polyadenylation events result in heterogeneous SAGE tags (Fig. 5.1).

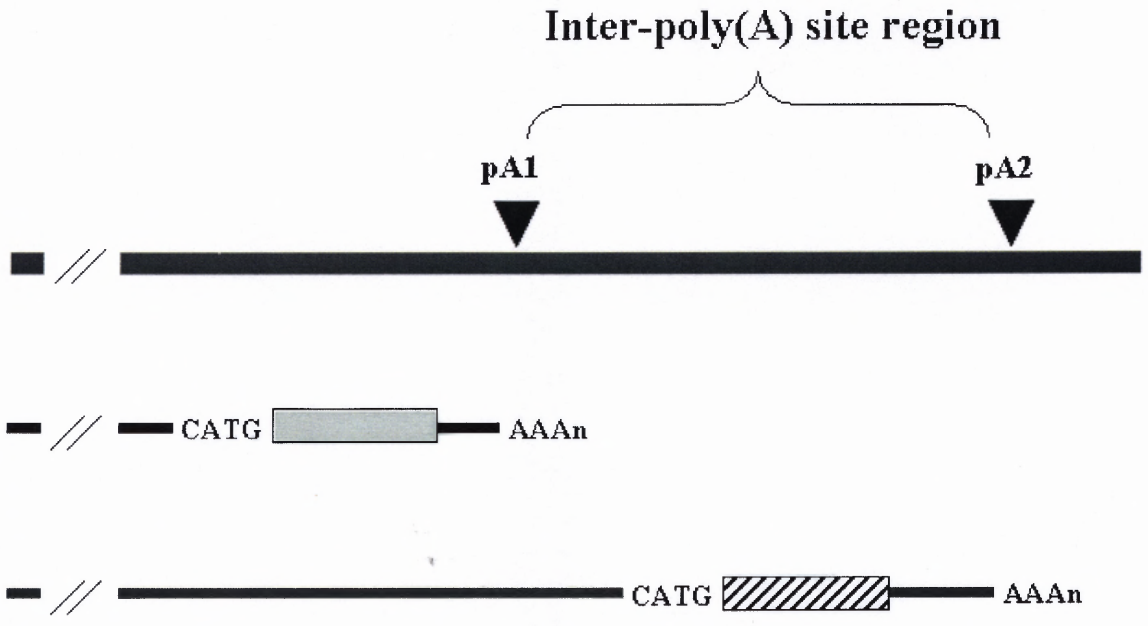


Figure 5.1 Schematic representation of alternate polyadenylation sites resulting in heterogeneous SAGE tags. mRNA is depicted as solid lines, pA1 and pA2 represent alternate polyadenylation sites. Filled bar and stripped bar represent two different heterogeneous SAGE tags, respectively. CATG represents anchoring enzyme NlaIII recognition site.

To further evaluate number of human genes that can be identified by heterogeneous SAGE tags, both *in silico* mapping of SAGE tags and experimental SAGE tags from multiple libraries were compared. The comprehensive and reliable SAGE tag-to-gene mapping was obtained from the NCBI SAGEmap project (Lash et al., 2000), where 294,122 distinct SAGE tags were mapped to a total of 38,971 UniGene in humans. Human SAGE data from 241 libraries were downloaded from NCBI GEO (Edgar et al., 2002), where 609,049 distinct SAGE tags were detected in at least one of the libraries. There are only 277,379 (46%) of these distinct tags can be mapped to known genes by

SAGEmap, whereas the rest 54% of SAGE tags detected in libraries cannot be matched to any known genes. As it has been shown experimentally that most of the unmatched tags are originated from novel transcripts instead of experimental errors (Chen et al. 2002), such a big portion of un-matched tags suggests that there are a large number of novel transcripts yet to be identified. A total of 35,547 out of 38,971 genes can be detected in at least one of the 241 human SAGE libraries, 20,906 (59%) of which having multiple heterogeneous SAGE tags. Finally, 16,743 (6%) of SAGEmap tags have never been detected in any of the 241 SAGE libraries, which map to 3,424 genes. These could be development- or tissue- specific transcripts that are not yet represented in the 241 SAGE libraries, or extremely low-level of transcripts that cannot be detected by the current methods.

5.3.2 Combining SAGE Data with EST Data

5.3.2.1 Mapping SAGE Tags to Poly(A) Sites

To combine SAGE data with poly(A) sites information derived from ESTs data, 407,169 human poly(A)/(T) tailed ESTs supporting poly(A) sites annotated in PolyA_DB (Chapter 1) were used to search for SAGE tags (see methods for details). SAGE tags can be mapped on 349,926 EST sequences (86%) where NlaIII recognition sites (CATG) can be located. This mapping covers 27,864 (95%) out of 29,283 poly(A) sites annotated in PolyA_DB (Table 5.1). A total of 37,234 distinct SAGE tags are found in this mapping procedure, 91% of which (33,914) can be detected in at least one of the available 241 SAGE library data.

Table 5.1 Mapping SAGE tags to human poly(A) sites.

	Homo Sapiens
Total Number of ESTs Searched	407,169
Number of ESTs with SAGE tags	349,926
Number of ESTs without SAGE tags	57,243
Total Number of poly(A) sites in PolyA_DB	29,283
Number of poly(A) sites with SAGE Tags Mapped	27,864
Number of poly(A) sites without SAGE Tags Mapped	1,419

5.3.2.2 SAGE Data Provides Quantitative Measurements of the Differential Selection of Alternative Polyadenylation Sites

To demonstrate that SAGE data can be used to study alternative poly(A) in combination with EST data based on the mapping effort described above, two genes were studied using SAGE data, PAP II (Gene ID: 10914) and CstF-77 (Gene ID: 1479). The mapping of SAGE tags to poly(A) sites may result in multiple tags corresponding to one poly(A) sites, there are four possible reasons for this: 1) Alternative splicing; 2) Sequencing error; and 3) single nucleotide polymorphism (SNP). The latter two cases usually yield SAGE tags different only at one base position, which makes them very different from multiple tags due to alternative splicing. Another important thing to consider is the uniqueness of a SAGE tag. 88% (32,702 out of 37,234) SAGE tags uniquely map to one poly(A) site, whereas other 12% can be mapped to more than one poly(A) sites. Table 5.2 summarize SAGE tag mapping information to poly(A) sites of PAP II and CstF77.

Three poly(A) sites of CstF-77 are found based on EST data, suggesting CstF-77 itself could be regulated by alternative polyadenylation. Because CstF-77 is a key component of the polyadenylation machinery, this prompts a possibility of a negative

auto-regulation mechanism. In fact, Drosophila ortholog of CstF-77, *suppressor of forked* [*su(f)*], is indeed regulated by such a mechanism (Audibert et al., 1998; Audibert and Simonelig, 1998). Two hypotheses can be proposed for the negative feedback auto-regulation: 1) since selection of upstream intronic poly(A) site of human CstF-77 results in a short form of transcript, it might be subjected to mRNA surveillance for degradation; 2) the selection of the upstream poly(A) site will result in a different protein product lacking certain domains and the regulation happens on the protein level. If the first hypothesis is true, the level of short form transcript is expected to be much lower than that of the long form transcripts. Therefore, a comparison of transcript levels of the short and long forms of CstF-77 will provide insights into the regulation mechanisms. Furthermore, since tissue-specific autoregulation of *su(f)* has been observed in Drosophila, it will be interesting to explore if the selection of different human CstF-77 poly(A) sites results in differential pattern across the 241 human SAGE libraries.

CstF-77 has five SAGE tags from 19 supporting ESTs mapped to 3 poly(A) sites. TTCCCAGNGA cannot be used to study expression pattern because it has an ambiguous base (N) and only one supporting EST (BG194969, Table 5.2). CTTCCTCTGC was also mapped to another poly(A) site, and thus cannot be used to infer expression pattern of usage of the third poly(A) site of CstF-77. Mapping of the rest three tags (TCAGGAGACG, GTTTTTGAGA, and GTTCTTGAGA) to the RefSeq mRNA of CstF-77 readily provides quantitative assessment of poly(A) site selections across different libraries (Figure 5.2). It worth to point out that GTTTTTGAGA and GTTCTTGAGA have only one base different from each other, which is probably a single nucleotide polymorphism (SNP) candidate. However, searching of 219 SNPs associated

with CstF-77 did not find this has been previously reported in NCBI dbSNP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Display&dopt=gene_snp&from_uid=1479).

SAGE data are presented in TPM (tag per million) format as a direct measure of mRNA expression levels. This is to normalize total SAGE-tag counts across libraries so that they are comparable (see method for details). Overall, differential selections of long and short forms of CstF-77 can be observed across 241 SAGE libraries (Figure 5.2 bottom). However, the total TPM values across 241 libraries of the long and short forms of human CstF-77 are comparable (TPM 2743.48 vs. TPM 1675.59), indicating that the auto-regulation of CstF-77 in humans might not be dependent on mRNA degradation of the short transcripts.

Another example was shown using PAP II (Figure 5.3), which has already been proven to be regulated via alternative polyadenylation (Zhao and Manley, 1996). Only tags mapped to exons presented in the RefSeq of PAP II (NM_032632) were displayed in the alignments.

Table 5.2 Annotation of SAGE tag mapped to poly(A) sites of CstF-77 and PAP II.

	Poly(A) site	SAGE tag	Supporting ESTs (tag position)	Other Information
CstF-77	p.1479.1	TTCCCAGNGA	BG194969 (28)	
		TCAGGAGACG	BM478846 (23), BQ184091 (457), BU685495 (579), BC059948 (23), BC009792 (151)	
	p.1479.2	GTTCTTGAGA	AI041674 (368), AI091381 (368), AI142702 (368), CB250473 (371), CB850814 (366), BG232033 (369), BM669021 (366), BU685894 (366), CA312681 (409), CA418384 (366), CA449643 (371), AA700556 (368), AA748501 (368), AA767348 (371), BC010533 (886)	
		GTTTTTGAGA	AI743827 (363), BQ446780 (366)	
	p.1479.3	CTTCCTCTGC	AI539467 (89), AI698588 (89)	p.153830.1
PAP II	p.10914.1	TTGATCGAAG	BF055552 (378), BQ575160 (343), BU626344 (382)	
	p.10914.2	TTGATCGAAG	CB132302 (90), AI310137 (512), AI420619 (513), AI497615 (511), BX367811 (196), CD368864 (512), BM969455 (512), BM970750 (512), BM977510 (512), BQ015247 (507), BQ574600 (512), BQ581604 (503), BU071418 (512), BU677628 (512), BU681961 (513), BU682027 (512), BC000927 (538)	
		GTATGTGTTA	BG611257 (655)	
	p.10914.3	TAGGCAGCCT	AA888232 (328	
	p.10914.4	TCCGTATTAC	AA923353 (217), CB055252 (219), AW205161 (217), BU621947 (219	
	p.10914.5	N/A	N/A	
	p.10914.6	ACTGGAGTTC	BG288697 (263), BM478941 (687), AA580599 (113)	ACTGGAGTT [C/T] allele is documented in dbSNP rs12147717
		ACTGGAGTTG	BF797555 (261)	
		ACTGGAGTTT	AI922188 (115), BX384856 (496), CD644337 (553), CD644443 (553), CD655093 (594), CD655865 (553), CD656209 (591), AW058288 (115), BM679335 (115), BQ008843 (114), AA236200 (118), CA415002 (111), AA748677 (113), AA937350 (113), NM_032632 (4397), AF002990 (289), BC036014 (4397)	

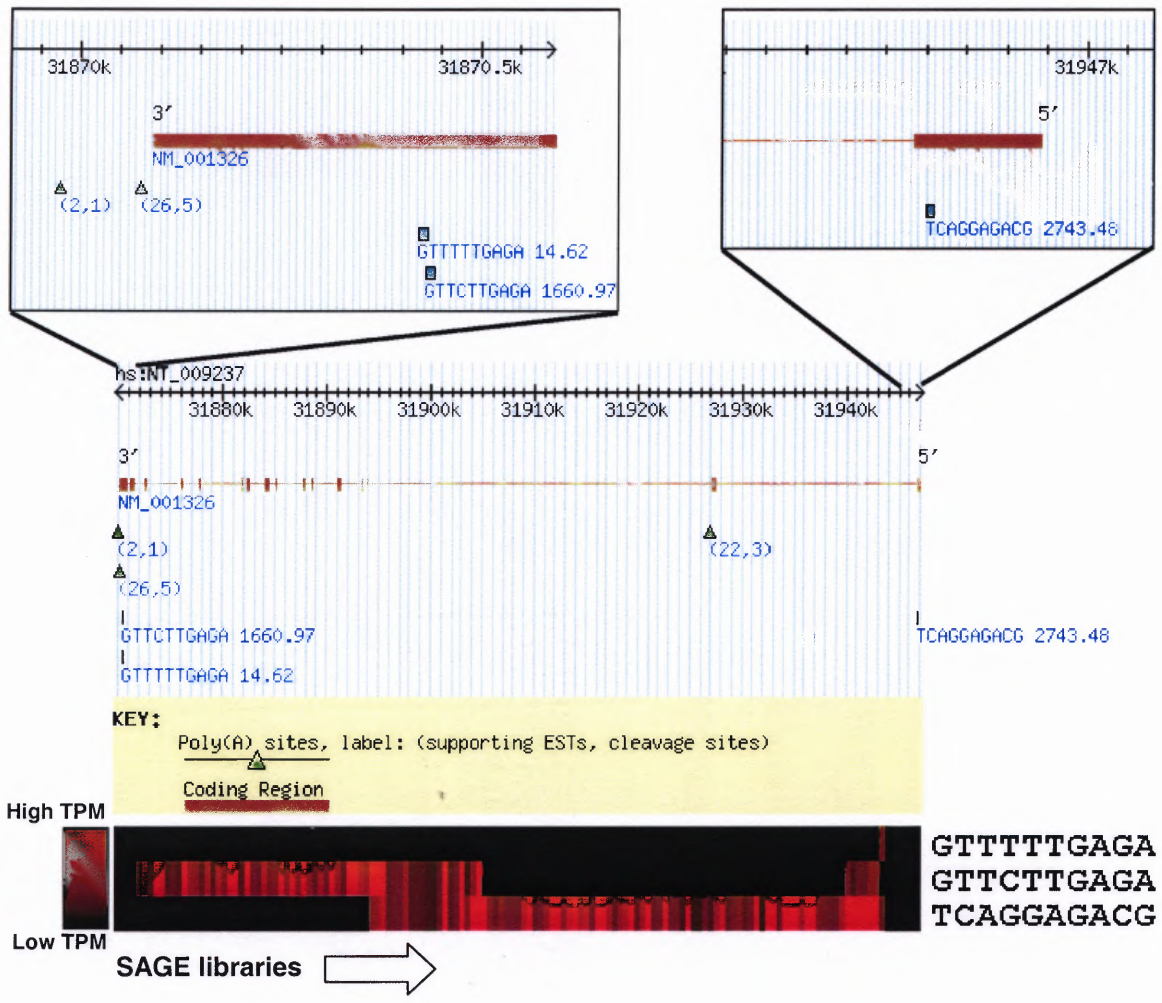


Figure 5.2 Mapping SAGE tags to Poly(A) sites of CstF-77. Shown in the middle is the overview of the gene structure, upstream and downstream regions containing poly(A) sites are shown in blown up image, numbers adjacent to SAGE tags are the sum of TPM (see method) across 241 human libraries. The bottom panel shows the TPM value of different tags across SAGE libraries; only libraries with detected tags were shown.

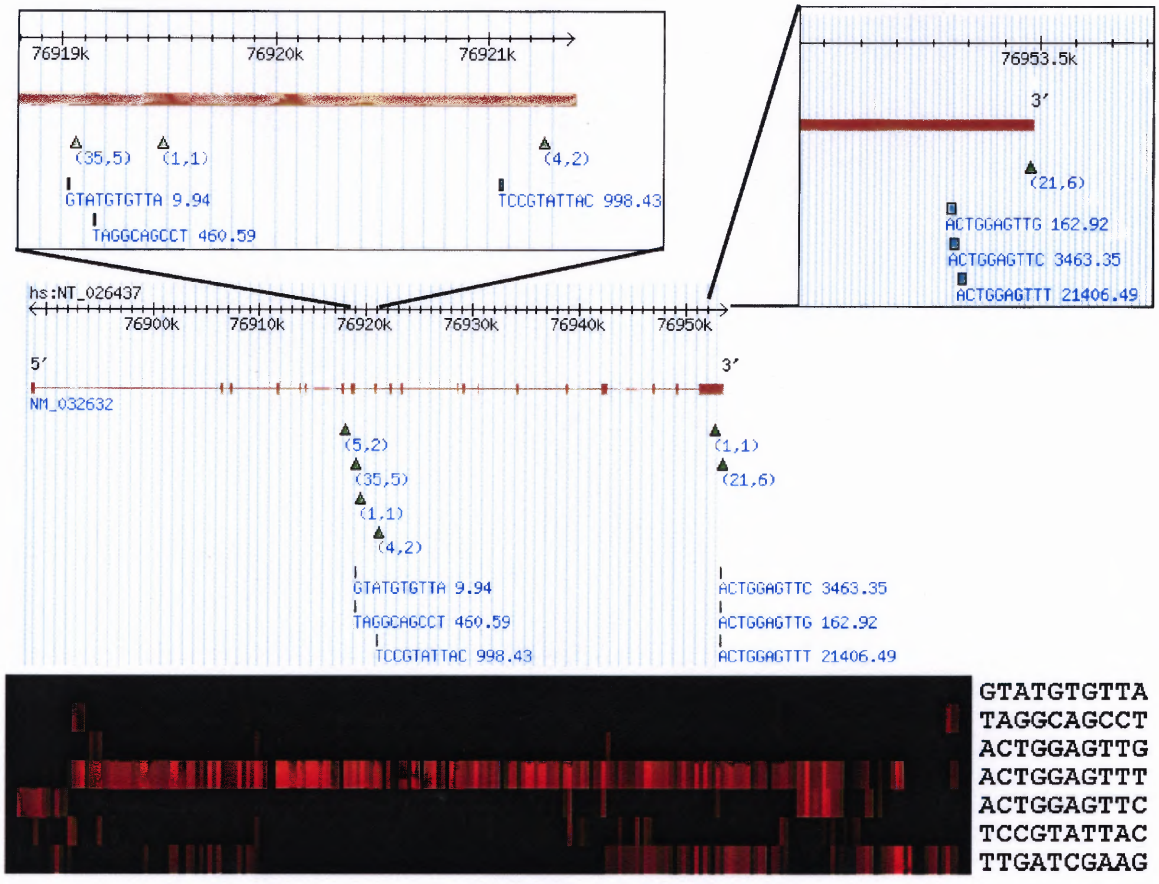


Figure 5.3 Mapping SAGE tags to Poly(A) sites of PAP II. See Figure 5.2 for descriptions.

5.3.3 Discussion

Although the conventional 14nt SAGE tags (10 nt after CATG) allow high-level unique assignments to ESTs, they do not perform well when considering duplicated genes or repeated sequences and putting in genomic context. Recently, a technique called longSAGE has been developed, where a different tagging enzyme (MmeI) is used to locate a much longer tag (21 nt including the tagging enzyme recognition site, unique 17 nt sequence) to represent each transcript (Saha et al., 2002). Theoretically >99.8% longSAGE tags only are expected to occur once in the human genome (Saha et al., 2002). Therefore, the uniqueness of tags representing transcripts is dramatically increased.

However, there are only two longSAGE datasets deposited in NCBI GEO database currently (one for humans and one for mice). Applying a similar strategy as described here will provide a more thorough systematic view of alternative polyadenylation when more data are available.

5.4 Materials and Methods

5.4.1 SAGE Data Analysis

SAGE data was downloaded from NCBI GEO (Edgar et al., 2002). SAGE tag counts are converted to TPM (tags per million) according to the total tag count in each library as follows:

$$TPM_i(j) = C_i(j) * 10^6 / T(j)$$

where $TPM_i(j)$ is TPM for SAGE tag i in SAGE library j, $C_i(j)$ is the absolute count for SAGE tag i in library j, and $T(j)$ is the total SAGE tag count in SAGE library j. NCBI reliable SAGE tag mapping set was downloaded, which is derived from best alignments of RefSeq, MGC, mRNA sources, and EST sequences to the draft human genomic sequence (ftp://ftp.ncbi.nih.gov/pub/sage/map/readme.txt, Lash et al 2000).

5.4.2 Mapping of SAGE Tags to Poly(A) Sites

EST sequences are downloaded from NCBI dbEST (March, 2004 release). A virtual SAGE tag set was first established by searching through all poly(A)-sites-supporting ESTs. A perl program is written to look for NlaIII recognition sites (CATG) in Poly(A)/(T) tailed ESTs supporting poly(A) sites in PolyA_DB. Whenever CATG is found, the 10 nt SAGE tag to the 3' of the mRNA was extracted. If the sequence to the 3'

is less than 10 nt, A's are added to bring the length of the tag to 10. The virtual SAGE tag set was then compared with SAGE tags detected in the set of 241 SAGE libraries for supporting expression level of transcripts resulted from different choice of poly(A) sites.

5.4.3 Visualization of SAGE Data

SAGE data is visualized by TreeView program (Eisen et al., 1998), graphic representations of alignments and positions of SAGE tags are produced using BioPerl (Stajich et al., 2002). To represent to gene structure, RefSeq sequences are used. Therefore, all SAGE tags are also mapped to RefSeq mRNA sequences. This could affect some SAGE tags that are mapped to the alternative spliced exons that are not represented in RefSeqs.

REFERENCES

Arhin, G. K., Boots, M., Bagga, P. S., Milcarek, C., and Wilusz, J. (2002). Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals. Nucleic Acids Res *30*, 1842-1850.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Ashiya, M., and Grabowski, P. J. (1997). A neuron-specific splicing switch mediated by an array of pre-mRNA repressor sites: evidence of a regulatory role for the polypyrimidine tract binding protein and a brain-specific PTB counterpart. Rna *3*, 996-1015.

Audibert, A., Juge, F., and Simonelig, M. (1998). The suppressor of forked protein of Drosophila, a homologue of the human 77K protein required for mRNA 3'-end formation, accumulates in mitotically-active cells. Mech Dev *72*, 53-63.

Audibert, A., and Simonelig, M. (1998). Autoregulation at the level of mRNA 3' end formation of the suppressor of forked gene of Drosophila melanogaster is conserved in Drosophila virilis. Proc Natl Acad Sci U S A *95*, 14302-14307.

Bai, C., and Tolias, P. P. (1996). Cleavage of RNA hairpins mediated by a developmentally regulated CCCH zinc finger protein. Mol Cell Biol *16*, 6661-6667.

Bai, C., and Tolias, P. P. (1998). Drosophila clipper/CPSF 30K is a post-transcriptionally regulated nuclear protein that binds RNA containing GC clusters. Nucleic Acids Res *26*, 1597-1604.

Ball, C. A., Awad, I. A., Demeter, J., Gollub, J., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Matese, J. C., Nitzberg, M., Wymore, F., *et al.* (2005). The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res *33 Database Issue*, D580-582.

Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res *10*, 1001-1010.

Beaudoing, E., and Gautheret, D. (2001). Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res *11*, 1520-1526.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Ser B *57*, 289-300.

Bienroth, S., Wahle, E., Suter-Crazzolara, C., and Keller, W. (1991). Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors. J Biol Chem *266*, 19768-19776.

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993). dbEST--database for "expressed sequence tags". Nat Genet *4*, 332-333.

Bonaldo, M. F., Lennon, G., and Soares, M. B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res *6*, 791-806.

Brackenridge, S., Ashe, H. L., Giacca, M., and Proudfoot, N. J. (1997). Transcription and polyadenylation in a short human intergenic region. Nucleic Acids Res *25*, 2326-2336.

Brackenridge, S., and Proudfoot, N. J. (2000). Recruitment of a basal polyadenylation factor by the upstream sequence element of the human lamin B2 polyadenylation signal. Mol Cell Biol *20*, 2660-2669.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., *et al.* (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res *31*, 68-71.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett *474*, 83-86.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. Nat Genet *30*, 29-30.

Brilliant, M. H., Sueoka, N., and Chikaraishi, D. M. (1984). Cloning of DNA corresponding to rare transcripts of rat brain: evidence of transcriptional and post-transcriptional control and of the existence of nonpolyadenylated transcripts. Mol Cell Biol *4*, 2187-2197.

Calvo, O., and Manley, J. L. (2001). Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination. Mol Cell *7*, 1013-1023.

Calvo, O., and Manley, J. L. (2003). Strange bedfellows: polyadenylation factors at the promoter. Genes Dev *17*, 1321-1327.

Carswell, S., and Alwine, J. C. (1989). Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences. Mol Cell Biol *9*, 4248-4258.

Castelo-Branco, P., Furger, A., Wollerton, M., Smith, C., Moreira, A., and Proudfoot, N. (2004). Polypyrimidine tract binding protein modulates efficiency of polyadenylation. Mol Cell Biol *24*, 4174-4183.

Chan, R. C., and Black, D. L. (1997). The polypyrimidine tract binding protein binds upstream of neural cell-specific c-src exon N1 to repress the splicing of the intron downstream. Mol Cell Biol *17*, 4667-4676.

Chen, F., MacDonald, C. C., and Wilusz, J. (1995). Cleavage site determinants in the mammalian polyadenylation signal. Nucleic Acids Res *23*, 2614-2620.

Chen, F., and Wilusz, J. (1998). Auxiliary downstream elements are required for efficient polyadenylation of mammalian pre-mRNAs. Nucleic Acids Res *26*, 2891-2898.

Chou, Z. F., Chen, F., and Wilusz, J. (1994). Sequence and position requirements for uridylate-rich downstream elements of polyadenylation signals. Nucleic Acids Res *22*, 2525-2531.

Colgan, D. F., and Manley, J. L. (1997). Mechanism and regulation of mRNA polyadenylation. Genes Dev *11*, 2755-2766.

Curtis, D., Lehmann, R., and Zamore, P. D. (1995). Translational regulation in development. Cell *81*, 171-178.

Dantonel, J. C., Murthy, K. G., Manley, J. L., and Tora, L. (1997). Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. Nature *389*, 399-402.

de Vries, H., Ruegsegger, U., Hubner, W., Friedlein, A., Langen, H., and Keller, W. (2000). Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. Embo J *19*, 5895-5904.

Denome, R. M., and Cole, C. N. (1988). Patterns of polyadenylation site selection in gene constructs containing multiple polyadenylation signals. Mol Cell Biol *8*, 4829-4839.

Dominski, Z., and Marzluff, W. F. (1999). Formation of the 3' end of histone mRNA. Gene *239*, 1-14.

Dominski, Z., Yang, X. C., Purdy, M., Wagner, E. J., and Marzluff, W. F. (2005). A CPSF-73 homologue is required for cell cycle progression but not cell growth and interacts with a protein having features of CPSF-100. Mol Cell Biol *25*, 1489-1500.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res *30*, 207-210.

Edwalds-Gilbert, G., Veraldi, K. L., and Milcarek, C. (1997). Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res *25*, 2547-2561.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*, 14863-14868.

Fairbrother, W. G., Yeh, R. F., Sharp, P. A., and Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007-1013.

Fong, N., and Bentley, D. L. (2001). Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. Genes Dev *15*, 1783-1795.

Fung, B. P., Brilliant, M. H., and Chikaraishi, D. M. (1991). Brain-specific polyA-transcripts are detected in polyA+ RNA: do complex polyA- brain RNAs really exist? J Neurosci *11*, 701-708.

Garcia-Blanco, M. A., Jamison, S. F., and Sharp, P. A. (1989). Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. Genes Dev *3*, 1874-1886.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., and Claverie, J. M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. Genome Res *8*, 524-530.

Ge, H., and Roeder, R. G. (1994). Purification, cloning, and characterization of a human coactivator, PC4, that mediates transcriptional activation of class II genes. Cell *78*, 513-523.

Gehring, N. H., Frede, U., Neu-Yilik, G., Hundsdoerfer, P., Vetter, B., Hentze, M. W., and Kulozik, A. E. (2001). Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. Nat Genet *28*, 389-392.

Gieselmann, V., Polten, A., Kreysing, J., and von Figura, K. (1989). Arylsulfatase A pseudodeficiency: loss of a polyadenylylation signal and N-glycosylation site. Proc Natl Acad Sci U S A *86*, 9436-9440.

Gil, A., and Proudfoot, N. J. (1987). Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. Cell *49*, 399-406.

Gilmartin, G. M., Fleming, E. S., Oetjen, J., and Graveley, B. R. (1995). CPSF recognition of an HIV-1 mRNA 3'-processing enhancer: multiple sequence contacts involved in poly(A) site definition. Genes Dev *9*, 72-83.

Graber, J. H., Cantor, C. R., Mohr, S. C., and Smith, T. F. (1999a). Genomic detection of new yeast pre-mRNA 3'-end-processing signals. Nucleic Acids Res *27*, 888-894.

Graber, J. H., Cantor, C. R., Mohr, S. C., and Smith, T. F. (1999b). In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. Proc Natl Acad Sci U S A *96*, 14055-14060.

Gunderson, S. I., Beyer, K., Martin, G., Keller, W., Boelens, W. C., and Mattaj, L. W. (1994). The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. Cell *76*, 531-541.

Gunderson, S. I., Polycarpou-Schwarz, M., and Mattaj, I. W. (1998). U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. Mol Cell *1*, 255-264.

Gunderson, S. I., Vagner, S., Polycarpou-Schwarz, M., and Mattaj, I. W. (1997). Involvement of the carboxyl terminus of vertebrate poly(A) polymerase in U1A autoregulation and in the coupling of splicing and polyadenylation. Genes Dev *11*, 761-773.

Guntaka, R. V. (1993). Transcription termination and polyadenylation in retroviruses. Microbiol Rev *57*, 511-521.

Hans, H., and Alwine, J. C. (2000). Functionally significant secondary structure of the simian virus 40 late polyadenylation signal. Mol Cell Biol *20*, 2926-2932.

He, X., Khan, A. U., Cheng, H., Pappas, D. L., Jr., Hampsey, M., and Moore, C. L. (2003). Functional interactions between the transcription and mRNA 3' end processing machineries mediated by Ssu72 and Sub1. Genes Dev *17*, 1030-1042.

Higgs, D. R., Goodbourn, S. E., Lamb, J., Clegg, J. B., Weatherall, D. J., and Proudfoot, N. J. (1983). Alpha-thalassaemia caused by a polyadenylation signal mutation. Nature *306*, 398-400.

Hirose, Y., and Manley, J. L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. Nature *395*, 93-96.

Hirose, Y., and Manley, J. L. (2000). RNA polymerase II and the integration of nuclear events. Genes Dev *14*, 1415-1429.

Hirose, Y., Tacke, R., and Manley, J. L. (1999). Phosphorylated RNA polymerase II stimulates pre-mRNA splicing. Genes Dev *13*, 1234-1239.

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141-2144.

Kan, Z., Rouchka, E. C., Gish, W. R., and States, D. J. (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res *11*, 889-900.

Kaufmann, I., Martin, G., Friedlein, A., Langen, H., and Keller, W. (2004). Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. Embo J *23*, 616-626.

Landry, J. R., Mager, D. L., and Wilhelm, B. T. (2003). Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet *19*, 640-648.

Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J., and Altschul, S. F. (2000). SAGEmap: a public gene expression resource. Genome Res *10*, 1051-1060.

Legendre, M., and Gautheret, D. (2003). Sequence determinants in human polyadenylation site selection. BMC Genomics *4*, 7.

Lewis, J. D., Gunderson, S. I., and Mattaj, I. W. (1995). The influence of 5' and 3' end structures on pre-mRNA metabolism. J Cell Sci Suppl *19*, 13-19.

Licatalosi, D. D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J. B., and Bentley, D. L. (2002). Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. Mol Cell *9*, 1101-1111.

Lin, C. L., Bristol, L. A., Jin, L., Dykes-Hoberg, M., Crawford, T., Clawson, L., and Rothstein, J. D. (1998). Aberrant RNA processing in a neurodegenerative disease: the cause for absent EAAT2, a glutamate transporter, in amyotrophic lateral sclerosis. Neuron *20*, 589-602.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol *14*, 1675-1680.

Lockhart, D. J., and Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. Nature *405*, 827-836.

Lou, H., Gagel, R. F., and Berget, S. M. (1996). An intron enhancer recognized by splicing factors activates polyadenylation. Genes Dev *10*, 208-219.

Lou, H., Neugebauer, K. M., Gagel, R. F., and Berget, S. M. (1998). Regulation of alternative polyadenylation by U1 snRNPs and SRp20. Mol Cell Biol *18*, 4977-4985.

Lutz, C. S., and Alwine, J. C. (1994). Direct interaction of the U1 snRNP-A protein with the upstream efficiency element of the SV40 late polyadenylation signal. Genes Dev *8*, 576-586.

Lutz, C. S., Murthy, K. G., Schek, N., O'Connor, J. P., Manley, J. L., and Alwine, J. C. (1996). Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. Genes Dev *10*, 325-337.

MacDonald, C. C., and Redondo, J. L. (2002). Reexamining the polyadenylation signal: were we wrong about AAUAAA? Mol Cell Endocrinol *190*, 1-8.

MacDonald, C. C., Wilusz, J., and Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Mol Cell Biol *14*, 6647-6654.

Markovtsov, V., Nikolic, J. M., Goldman, J. A., Turck, C. W., Chou, M. Y., and Black, D. L. (2000). Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. Mol Cell Biol *20*, 7463-7479.

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S. D., Wickens, M., and Bentley, D. L. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature *385*, 357-361.

McDevitt, M. A., Hart, R. P., Wong, W. W., and Nevins, J. R. (1986). Sequences capable of restoring poly(A) site function define two distinct downstream elements. Embo J *5*, 2907-2913.

McDevitt, M. A., Imperiale, M. J., Ali, H., and Nevins, J. R. (1984). Requirement of a downstream sequence for generation of a poly(A) addition site. Cell *37*, 993-999.

Millevoi, S., Geraghty, F., Idowu, B., Tam, J. L., Antoniou, M., and Vagner, S. (2002). A novel function for the U2AF 65 splicing factor in promoting pre-mRNA 3'-end processing. EMBO Rep *3*, 869-874.

Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res *29*, 2850-2859.

Moore, C. L., Chen, J., and Whoriskey, J. (1988). Two proteins crosslinked to RNA containing the adenovirus L3 poly(A) site require the AAUAAA sequence for binding. Embo J *7*, 3159-3169.

Moreira, A., Takagaki, Y., Brackenridge, S., Wollerton, M., Manley, J. L., and Proudfoot, N. J. (1998). The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3' end formation by two distinct mechanisms. Genes Dev *12*, 2522-2534.

Moreira, A., Wollerton, M., Monks, J., and Proudfoot, N. J. (1995). Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. Embo J *14*, 3809-3819.

Murthy, K. G., and Manley, J. L. (1992). Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. J Biol Chem *267*, 14804-14811.

Murthy, K. G., and Manley, J. L. (1995). The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. Genes Dev *9*, 2672-2683.

Natalizio, B. J., Muniz, L. C., Arhin, G. K., Wilusz, J., and Lutz, C. S. (2002). Upstream elements present in the 3'-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals. J Biol Chem *277*, 42733-42740.

Orkin, S. H., Cheng, T. C., Antonarakis, S. E., and Kazazian, H. H., Jr. (1985). Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. Embo J *4*, 453-456.

Orphanides, G., and Reinberg, D. (2002). A unified theory of gene expression. Cell *108*, 439-451.

Pauws, E., van Kampen, A. H., van de Graaf, S. A., de Vijlder, J. J., and Ris-Stalpers, C. (2001). Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. Nucleic Acids Res *29*, 1690-1694.

Perez Canadillas, J. M., and Varani, G. (2003). Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. Embo J *22*, 2821-2830.

Peterson, M. L., Gimmi, E. R., and Perry, R. P. (1991). The developmentally regulated shift from membrane to secreted mu mRNA production is accompanied by an increase in cleavage-polyadenylation efficiency but no measurable change in splicing efficiency. Mol Cell Biol *11*, 2324-2327.

Phillips, C., and Gunderson, S. (2003). Sequences adjacent to the 5' splice site control U1A binding upstream of the IgM heavy chain secretory poly(A) site. J Biol Chem *278*, 22102-22111.

Phillips, C., Jung, S., and Gunderson, S. I. (2001). Regulation of nuclear poly(A) addition controls the expression of immunoglobulin M secretory mRNA. Embo J *20*, 6443-6452.

Phillips, C., Kyriakopoulou, C. B., and Virtanen, A. (1999). Identification of a stem-loop structure important for polyadenylation at the murine IgM secretory poly(A) site. Nucleic Acids Res *27*, 429-438.

Phillips, C., Pachikara, N., and Gunderson, S. I. (2004). U1A inhibits cleavage at the immunoglobulin M heavy-chain secretory poly(A) site by binding between the two downstream GU-rich regions. Mol Cell Biol *24*, 6162-6171.

Proudfoot, N. (1996). Ending the message is not so simple. Cell *87*, 779-781.

Proudfoot, N. (2000). Connecting transcription to messenger RNA processing. Trends Biochem Sci *25*, 290-293.

Proudfoot, N. (2004). New perspectives on connecting messenger RNA 3' end formation to transcription. Curr Opin Cell Biol *16*, 272-278.

Proudfoot, N. J., and Brownlee, G. G. (1976). 3' non-coding region sequences in eukaryotic messenger RNA. Nature *263*, 211-214.

Proudfoot, N. J., Furger, A., and Dye, M. J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501-512.

Pruitt, K. D., and Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res *29*, 137-140.

Resch, A., Xing, Y., Alekseyenko, A., Modrek, B., and Lee, C. (2004). Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. Nucleic Acids Res *32*, 1261-1269.

Richter, J. D. (1999). Cytoplasmic polyadenylation in development and beyond. Microbiol Mol Biol Rev *63*, 446-456.

Ross, J. (1995). mRNA stability in mammalian cells. Microbiol Rev *59*, 423-450.

Ryan, K., Calvo, O., and Manley, J. L. (2004). Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. Rna *10*, 565-573.

Ryan, K., Murthy, K. G., Kaneko, S., and Manley, J. L. (2002). Requirements of the RNA polymerase II C-terminal domain for reconstituting pre-mRNA 3' cleavage. Mol Cell Biol *22*, 1684-1692.

Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2002). Using the transcriptome to annotate the genome. Nat Biotechnol *20*, 508-512.

Schek, N., Cooke, C., and Alwine, J. C. (1992). Definition of the upstream efficiency element of the simian virus 40 late polyadenylation signal by using in vitro analyses. Mol Cell Biol *12*, 5386-5393.

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res *18*, 6097-6100.

Sheets, M. D., Ogg, S. C., and Wickens, M. P. (1990). Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. Nucleic Acids Res *18*, 5799-5805.

Snider, B. J., and Morrison-Bogorad, M. (1992). Brain non-adenylated mRNAs. Brain Res Brain Res Rev *17*, 263-282.

Sorek, R., Shamir, R., and Ast, G. (2004). How prevalent is functional alternative splicing in the human genome? Trends Genet *20*, 68-71.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., *et al.* (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res *12*, 1611-1618.

Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., *et al.* (2002). Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A *99*, 4465-4470.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A *101*, 6062-6067.

Sugnet, C. W., Kent, W. J., Ares, M., Jr., and Haussler, D. (2004). Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac Symp Biocomput, 66-77.

Tabaska, J. E., and Zhang, M. Q. (1999). Detection of polyadenylation signals in human DNA sequences. Gene *231*, 77-86.

Takagaki, Y., and Manley, J. L. (1992). A human polyadenylation factor is a G protein beta-subunit homologue. J Biol Chem *267*, 23471-23474.

Takagaki, Y., and Manley, J. L. (1994). A polyadenylation factor subunit is the human homologue of the Drosophila suppressor of forked protein. Nature *372*, 471-474.

Takagaki, Y., and Manley, J. L. (1997). RNA recognition by the human polyadenylation factor CstF. Mol Cell Biol *17*, 3907-3914.

Takagaki, Y., and Manley, J. L. (1998). Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. Mol Cell *2*, 761-771.

Takagaki, Y., and Manley, J. L. (2000). Complex protein interactions within the human polyadenylation machinery identify a novel component. Mol Cell Biol *20*, 1515-1525.

Takagaki, Y., Ryner, L. C., and Manley, J. L. (1989). Four factors are required for 3'-end cleavage of pre-mRNAs. Genes Dev *3*, 1711-1724.

Takagaki, Y., Seipelt, R. L., Peterson, M. L., and Manley, J. L. (1996). The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. Cell *87*, 941-952.

Thanaraj, T. A., Clark, F., and Muilu, J. (2003). Conservation of human alternative splice events in mouse. Nucleic Acids Res *31*, 2544-2552.

Tian, B., Hu, J., Zhang, H., and Lutz, C. S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res *33*, 201-212.

Trinklein, N. D., Aldred, S. J., Saldanha, A. J., and Myers, R. M. (2003). Identification and functional analysis of human transcriptional promoters. Genome Res *13*, 308-312.

Valsamakis, A., Schek, N., and Alwine, J. C. (1992). Elements upstream of the AAUAAA within the human immunodeficiency virus polyadenylation signal are required for efficient polyadenylation in vitro. Mol Cell Biol *12*, 3699-3705.

Van Ness, J., Maxwell, I. H., and Hahn, W. E. (1979). Complex population of nonpolyadenylated messenger RNA in mouse brain. Cell *18*, 1341-1349.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. Science *270*, 484-487.

Veraldi, K. L., Arhin, G. K., Martincic, K., Chung-Ganster, L. H., Wilusz, J., and Milcarek, C. (2001). hnRNP F influences binding of a 64-kilodalton subunit of cleavage stimulation factor to mRNA precursors in mouse B cells. Mol Cell Biol *21*, 1228-1238.

Wahle, E., and Ruegsegger, U. (1999). 3'-End processing of pre-mRNA in eukaryotes. FEMS Microbiol Rev *23*, 277-295.

Wallace, A. M., Dass, B., Ravnik, S. E., Tonk, V., Jenkins, N. A., Gilbert, D. J., Copeland, N. G., and MacDonald, C. C. (1999). Two distinct forms of the 64,000 Mr protein of the cleavage stimulation factor are expressed in mouse male germ cells. Proc Natl Acad Sci U S A *96*, 6763-6768.

Wang, H., Hubbell, E., Hu, J. S., Mei, G., Cline, M., Lu, G., Clark, T., Siani-Rose, M. A., Ares, M., Kulp, D. C., and Haussler, D. (2003). Gene structure-based splice variant deconvolution using a microarray platform. Bioinformatics *19 Suppl 1*, i315-322.

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A., and Wagner, L. (2003). Database resources of the National Center for Biotechnology. Nucleic Acids Res *31*, 28-33.

Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A., and Smith, C. W. (2004). Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. Mol Cell *13*, 91-100.

Xing, H., Mayhew, C. N., Cullen, K. E., Park-Sarge, O. K., and Sarge, K. D. (2004). HSF1 modulation of Hsp70 mRNA polyadenylation via interaction with symplekin. J Biol Chem *279*, 10551-10555.

Xu, A., and Chen, K. Y. (2001). Hypusine is required for a sequence-specific interaction of eukaryotic initiation factor 5A with postsystematic evolution of ligands by exponential enrichment RNA. J Biol Chem *276*, 2555-2561.

Xu, A., Jao, D. L., and Chen, K. Y. (2004). Identification of mRNA that binds to eukaryotic initiation factor 5A by affinity co-purification and differential display. Biochem J *384*, 585-590.

Xu, Q., and Lee, C. (2003). Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. Nucleic Acids Res *31*, 5635-5643.

Xu, Q., Modrek, B., and Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res *30*, 3754-3766.

Yeo, G., Holste, D., Kreiman, G., and Burge, C. B. (2004). Variation in alternative splicing across human tissues. Genome Biol *5*, R74.

Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T., and Burge, C. B. (2005). Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A *102*, 2850-2855.

Zarudnaya, M. I., Kolomiets, I. M., and Hovorun, D. M. (2002). What nuclease cleaves pre-mRNA in the process of polyadenylation? IUBMB Life *54*, 27-31.

Zarudnaya, M. I., Kolomiets, I. M., Potyahaylo, A. L., and Hovorun, D. M. (2003). Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. Nucleic Acids Res *31*, 1375-1386.

Zhang, H., Hu, J., Recce, M., and Tian, B. (2005). PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res *33 Database Issue*, D116-120.

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev *63*, 405-445.

Zhao, W., and Manley, J. L. (1996). Complex alternative RNA processing generates an unexpected diversity of poly(A) polymerase isoforms. Mol Cell Biol *16*, 2378-2386.